

**Q1**

1. The covariance matrix is:

| **Covariance Matrix, DF = 92** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
|  | **X1** | **X2** | **X3** | **X4** | **X5** | **X6** |
| **X1** | 4.675178369 | 0.912885960 | 0.544211676 | 0.279335921 | 1.036604683 | 0.160167368 |
| **X2** | 0.912885960 | 0.623740949 | 0.101289339 | 0.124223943 | 0.385589563 | -0.019368864 |
| **X3** | 0.544211676 | 0.101289339 | 0.512917485 | 0.072324371 | 0.306767731 | 0.077622370 |
| **X4** | 0.279335921 | 0.124223943 | 0.072324371 | 0.110145979 | 0.216745755 | 0.007400421 |
| **X5** | 1.036604683 | 0.385589563 | 0.306767731 | 0.216745755 | 0.846696893 | -0.031518583 |
| **X6** | 0.160167368 | -0.019368864 | 0.077622370 | 0.007400421 | -0.031518583 | 0.864890136 |

The correlation matrix is:

| **Pearson Correlation Coefficients, N = 93  Prob > |r| under H0: Rho=0** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
|  | **X1** | **X2** | **X3** | **X4** | **X5** | **X6** |
| **X1** | |  | | --- | | 1.00000 | |  | | |  | | --- | | 0.53458 | | <.0001 | | |  | | --- | | 0.35143 | | 0.0006 | | |  | | --- | | 0.38926 | | 0.0001 | | |  | | --- | | 0.52101 | | <.0001 | | |  | | --- | | 0.07965 | | 0.4479 | |
| **X2** | |  | | --- | | 0.53458 | | <.0001 | | |  | | --- | | 1.00000 | |  | | |  | | --- | | 0.17908 | | 0.0859 | | |  | | --- | | 0.47394 | | <.0001 | | |  | | --- | | 0.53059 | | <.0001 | | |  | | --- | | -0.02637 | | 0.8019 | |
| **X3** | |  | | --- | | 0.35143 | | 0.0006 | | |  | | --- | | 0.17908 | | 0.0859 | | |  | | --- | | 1.00000 | |  | | |  | | --- | | 0.30428 | | 0.0030 | | |  | | --- | | 0.46550 | | <.0001 | | |  | | --- | | 0.11654 | | 0.2659 | |
| **X4** | |  | | --- | | 0.38926 | | 0.0001 | | |  | | --- | | 0.47394 | | <.0001 | | |  | | --- | | 0.30428 | | 0.0030 | | |  | | --- | | 1.00000 | |  | | |  | | --- | | 0.70974 | | <.0001 | | |  | | --- | | 0.02398 | | 0.8195 | |
| **X5** | |  | | --- | | 0.52101 | | <.0001 | | |  | | --- | | 0.53059 | | <.0001 | | |  | | --- | | 0.46550 | | <.0001 | | |  | | --- | | 0.70974 | | <.0001 | | |  | | --- | | 1.00000 | |  | | |  | | --- | | -0.03683 | | 0.7260 | |
| **X6** | |  | | --- | | 0.07965 | | 0.4479 | | |  | | --- | | -0.02637 | | 0.8019 | | |  | | --- | | 0.11654 | | 0.2659 | | |  | | --- | | 0.02398 | | 0.8195 | | |  | | --- | | -0.03683 | | 0.7260 | | |  | | --- | | 1.00000 | |  | |

1. We will use the correlation matrix R with the following justification: Covariance is relevant when the following are true:

* The results of principal component analysis depend on the scales at which the variables are measured.

>> Here the scales appear to be similar

* Variables with the highest sample variances will tend to be emphasized in the first few principal components.

>> We are not looking to over or under emphasize any particular component. Therefore correlation R will be more relevant

* Principal component analysis using the covariance function should only be considered if all of the variables have the same units of measurement.

>> We don’t know the units of the measurements from the question.

So in summary based on the 2nd point we choose to use R for the Principal component analysis.

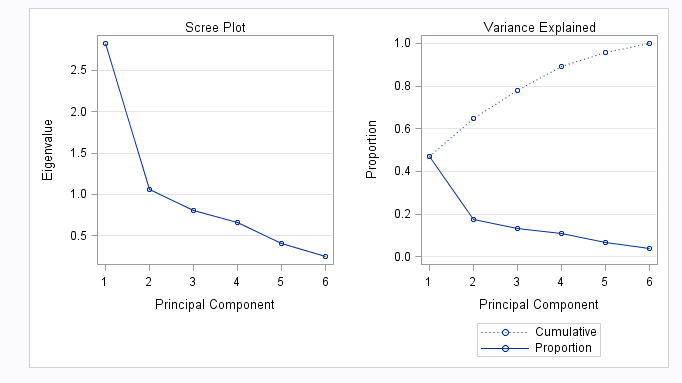
The table for eigen values is:

| **Eigenvalues of the Correlation Matrix** | | | | |
| --- | --- | --- | --- | --- |
|  | **Eigenvalue** | **Difference** | **Proportion** | **Cumulative** |
| **1** | 2.82524587 | 1.76489103 | 0.4709 | 0.4709 |
| **2** | 1.06035484 | 0.25569964 | 0.1767 | 0.6476 |
| **3** | 0.80465519 | 0.14607280 | 0.1341 | 0.7817 |
| **4** | 0.65858239 | 0.25387216 | 0.1098 | 0.8915 |
| **5** | 0.40471023 | 0.15825876 | 0.0675 | 0.9589 |
| **6** | 0.24645148 |  | 0.0411 | 1.0000 |

| **Eigenvectors** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Prin1** | **Prin2** | **Prin3** | **Prin4** | **Prin5** | **Prin6** |
| **X1** | 0.444503 | 0.067926 | 0.238098 | -.608552 | -.594830 | 0.130208 |
| **X2** | 0.437105 | -.210714 | 0.489278 | -.183126 | 0.698670 | 0.058910 |
| **X3** | 0.341981 | 0.360805 | -.733236 | -.267827 | 0.321219 | 0.200802 |
| **X4** | 0.469050 | -.090012 | 0.002962 | 0.648415 | -.199778 | 0.558146 |
| **X5** | 0.522938 | -.089193 | -.156486 | 0.252166 | -.112102 | -.786090 |
| **X6** | 0.030527 | 0.897077 | 0.376544 | 0.200918 | 0.048765 | -.098939 |

1. The first principal component explains about 47% of the variation. Furthermore, the first three principal components explain 78%, while the first four principal components explain 89% of the variation.

Since no cutoff is provided, lets look at the scree plot:



The only sharp drop that is noticeable in this case is after the first component. One might, based on this, select only one component. However, one component is probably too few, particularly because we have only explained 47% of the variation.

If we retain the first three components, we will explain 78% of the variance.

1. Assuming we retain the first three components, that will explain 78% of the variance.

| **Pearson Correlation Coefficients, N = 93  Prob > |r| under H0: Rho=0** | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Prin1** | **Prin2** | **Prin3** | **X1** | **X2** | **X3** | **X4** | **X5** | **X6** |
| **Prin1** | |  | | --- | | 1.00000 | |  | | |  | | --- | | 0.00000 | | 1.0000 | | |  | | --- | | 0.00000 | | 1.0000 | | |  | | --- | | 0.74714 | | <.0001 | | |  | | --- | | 0.73471 | | <.0001 | | |  | | --- | | 0.57482 | | <.0001 | | |  | | --- | | 0.78840 | | <.0001 | | |  | | --- | | 0.87898 | | <.0001 | | |  | | --- | | 0.05131 | | 0.6252 | |
| **Prin2** | |  | | --- | | 0.00000 | | 1.0000 | | |  | | --- | | 1.00000 | |  | | |  | | --- | | 0.00000 | | 1.0000 | | |  | | --- | | 0.06995 | | 0.5053 | | |  | | --- | | -0.21698 | | 0.0367 | | |  | | --- | | 0.37153 | | 0.0002 | | |  | | --- | | -0.09269 | | 0.3769 | | |  | | --- | | -0.09185 | | 0.3812 | | |  | | --- | | 0.92375 | | <.0001 | |
| **Prin3** | |  | | --- | | 0.00000 | | 1.0000 | | |  | | --- | | 0.00000 | | 1.0000 | | |  | | --- | | 1.00000 | |  | | |  | | --- | | 0.21358 | | 0.0398 | | |  | | --- | | 0.43889 | | <.0001 | | |  | | --- | | -0.65773 | | <.0001 | | |  | | --- | | 0.00266 | | 0.9798 | | |  | | --- | | -0.14037 | | 0.1796 | | |  | | --- | | 0.33777 | | 0.0009 | |
| **X1** | |  | | --- | | 0.74714 | | <.0001 | | |  | | --- | | 0.06995 | | 0.5053 | | |  | | --- | | 0.21358 | | 0.0398 | | |  | | --- | | 1.00000 | |  | | |  | | --- | | 0.53458 | | <.0001 | | |  | | --- | | 0.35143 | | 0.0006 | | |  | | --- | | 0.38926 | | 0.0001 | | |  | | --- | | 0.52101 | | <.0001 | | |  | | --- | | 0.07965 | | 0.4479 | |
| **X2** | |  | | --- | | 0.73471 | | <.0001 | | |  | | --- | | -0.21698 | | 0.0367 | | |  | | --- | | 0.43889 | | <.0001 | | |  | | --- | | 0.53458 | | <.0001 | | |  | | --- | | 1.00000 | |  | | |  | | --- | | 0.17908 | | 0.0859 | | |  | | --- | | 0.47394 | | <.0001 | | |  | | --- | | 0.53059 | | <.0001 | | |  | | --- | | -0.02637 | | 0.8019 | |
| **X3** | |  | | --- | | 0.57482 | | <.0001 | | |  | | --- | | 0.37153 | | 0.0002 | | |  | | --- | | -0.65773 | | <.0001 | | |  | | --- | | 0.35143 | | 0.0006 | | |  | | --- | | 0.17908 | | 0.0859 | | |  | | --- | | 1.00000 | |  | | |  | | --- | | 0.30428 | | 0.0030 | | |  | | --- | | 0.46550 | | <.0001 | | |  | | --- | | 0.11654 | | 0.2659 | |
| **X4** | |  | | --- | | 0.78840 | | <.0001 | | |  | | --- | | -0.09269 | | 0.3769 | | |  | | --- | | 0.00266 | | 0.9798 | | |  | | --- | | 0.38926 | | 0.0001 | | |  | | --- | | 0.47394 | | <.0001 | | |  | | --- | | 0.30428 | | 0.0030 | | |  | | --- | | 1.00000 | |  | | |  | | --- | | 0.70974 | | <.0001 | | |  | | --- | | 0.02398 | | 0.8195 | |
| **X5** | |  | | --- | | 0.87898 | | <.0001 | | |  | | --- | | -0.09185 | | 0.3812 | | |  | | --- | | -0.14037 | | 0.1796 | | |  | | --- | | 0.52101 | | <.0001 | | |  | | --- | | 0.53059 | | <.0001 | | |  | | --- | | 0.46550 | | <.0001 | | |  | | --- | | 0.70974 | | <.0001 | | |  | | --- | | 1.00000 | |  | | |  | | --- | | -0.03683 | | 0.7260 | |
| **X6** | |  | | --- | | 0.05131 | | 0.6252 | | |  | | --- | | 0.92375 | | <.0001 | | |  | | --- | | 0.33777 | | 0.0009 | | |  | | --- | | 0.07965 | | 0.4479 | | |  | | --- | | -0.02637 | | 0.8019 | | |  | | --- | | 0.11654 | | 0.2659 | | |  | | --- | | 0.02398 | | 0.8195 | | |  | | --- | | -0.03683 | | 0.7260 | | |  | | --- | | 1.00000 | |  | |

**First Principal Component Analysis - PCA1**

This seems to be most highly positively co-related to X1, X2, X4 and X5

**Second Principal Component Analysis – PCA2**

This seems to be most highly positively co-related to X1 and X6.

**Third Principal Component Analysis – PCA3**

This seems to be most highly negatively co-related to X3

**Q2**

****

1. Principal component with covariance

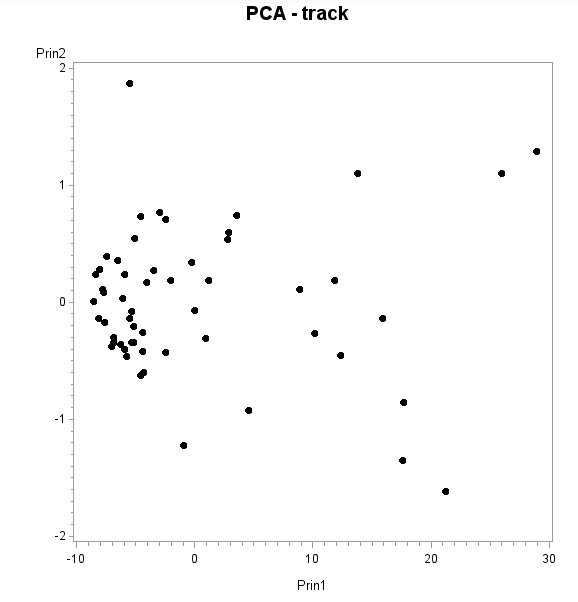


1. We get the following table:

| **Eigenvalues of the Covariance Matrix** | | | | |
| --- | --- | --- | --- | --- |
|  | **Eigenvalue** | **Difference** | **Proportion** | **Cumulative** |
| **1** | 88.6408123 | 88.2356498 | 0.9951 | 0.9951 |
| **2** | 0.4051626 | 0.3781763 | 0.0045 | 0.9997 |
| **3** | 0.0269863 | 0.0237128 | 0.0003 | 1.0000 |
| **4** | 0.0032735 | 0.0026772 | 0.0000 | 1.0000 |
| **5** | 0.0005963 | 0.0004590 | 0.0000 | 1.0000 |
| **6** | 0.0001373 | 0.0001078 | 0.0000 | 1.0000 |
| **7** | 0.0000295 | 0.0000260 | 0.0000 | 1.0000 |
| **8** | 0.0000035 |  | 0.0000 | 1.0000 |

It is clear from this table that only first component is required to explain more than 90% of the total variation.

Scatter plot:



1. We will refer to the table:

| **Eigenvectors** | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Prin1** | **Prin2** | **Prin3** | **Prin4** | **Prin5** | **Prin6** | **Prin7** | **Prin8** |
| **d100** | 0.000327 | 0.003927 | 0.000543 | 0.034157 | 0.070719 | 0.182717 | 0.382677 | 0.902214 |
| **d200** | 0.000687 | 0.006884 | 0.004641 | 0.067298 | 0.102726 | 0.279136 | 0.850732 | -.428004 |
| **d400** | 0.001832 | 0.013978 | 0.005671 | 0.132908 | 0.271722 | 0.881239 | -.359062 | -.052568 |
| **d800** | 0.005486 | 0.032506 | 0.021005 | 0.388687 | 0.857186 | -.334542 | -.026427 | -.003100 |
| **d1500** | 0.014415 | 0.086740 | 0.051904 | 0.901845 | -.419318 | -.013653 | -.013751 | 0.006908 |
| **d5000** | 0.079679 | 0.372494 | 0.920340 | -.088669 | 0.003055 | 0.000153 | -.000874 | 0.001253 |
| **d10000** | 0.181951 | 0.901522 | -.387002 | -.066050 | 0.002866 | -.003116 | 0.000118 | -.000900 |
| **marathon** | 0.979951 | -.199166 | -.003871 | 0.003724 | -.000014 | 0.000736 | 0.000346 | 0.000078 |

Y1 = 0.000327 \* d100 + 0.000687 \* d200 + 0.001832 \* d400 + 0.005486 \* d800 + 0.014415 \* d1500 + 0.079679 \* d5000 + 0.181951 \* d10000 + 0.979951 \* marathon

**Interpretation:** This component has the largest contribution from the value of marathon variable followed by 10K. This indicates that these two variables have the highest amount of variance and therefore are asserting the most influence.

We can validate this by looking at the following table:

| **Variable** | **N** | **Mean** | **Std Dev** | **Minimum** | **Maximum** |
| --- | --- | --- | --- | --- | --- |
| |  | | --- | | **d100** | | **d200** | | **d400** | | **d800** | | **d1500** | | **d5000** | | **d10000** | | **marathon** | | |  | | --- | | 55 | | 55 | | 55 | | 55 | | 55 | | 55 | | 55 | | 55 | | |  | | --- | | 0.1745182 | | 0.3490061 | | 0.7739788 | | 1.7932727 | | 3.6981818 | | 13.8458182 | | 28.9890909 | | 136.6240000 | | |  | | --- | | 0.0058572 | | 0.0107441 | | 0.0242836 | | 0.0636848 | | 0.1559094 | | 0.8011605 | | 1.8077316 | | 9.2270335 | | |  | | --- | | 0.1655000 | | 0.3286667 | | 0.7310000 | | 1.7000000 | | 3.5100000 | | 13.0100000 | | 27.3800000 | | 128.2200000 | | |  | | --- | | 0.2030000 | | 0.3866667 | | 0.8823333 | | 2.0200000 | | 4.2400000 | | 16.7000000 | | 35.3800000 | | 164.7000000 | |

It is clear that the variance of marathon is far greater than the variance of any of the other variables.

1. The smallest and largest values for Principal Component 1 are:

|  |  |  |
| --- | --- | --- |
| Country | Prin1 | Prin1 with translated variable \* |
| Usa | 131.8495 | -8.579 |
| |  | | --- | | Cookis | | 169.3551 | 28.926 |

\*uses the difference between the variables and their sample means

1. Principal component with covariance

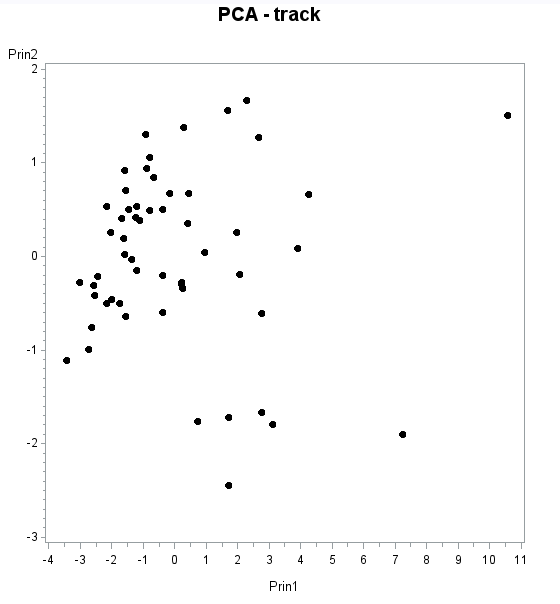


1. We get the following table:

| **Eigenvalues of the Correlation Matrix** | | | | |
| --- | --- | --- | --- | --- |
|  | **Eigenvalue** | **Difference** | **Proportion** | **Cumulative** |
| **1** | 6.62214613 | 5.74452784 | 0.8278 | 0.8278 |
| **2** | 0.87761829 | 0.71829715 | 0.1097 | 0.9375 |
| **3** | 0.15932114 | 0.03527176 | 0.0199 | 0.9574 |
| **4** | 0.12404939 | 0.04416911 | 0.0155 | 0.9729 |
| **5** | 0.07988027 | 0.01191512 | 0.0100 | 0.9829 |
| **6** | 0.06796515 | 0.02154562 | 0.0085 | 0.9914 |
| **7** | 0.04641953 | 0.02381943 | 0.0058 | 0.9972 |
| **8** | 0.02260010 |  | 0.0028 | 1.0000 |

It is clear from this table that we need first two components to explain more than 90% of the total variation.

Scatter plot:



1. We will refer to the table:

| **Eigenvectors** | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Prin1** | **Prin2** | **Prin3** | **Prin4** | **Prin5** | **Prin6** | **Prin7** | **Prin8** |
| **d100** | 0.317556 | 0.566878 | 0.332262 | 0.127628 | 0.262555 | -.593704 | 0.136241 | 0.105542 |
| **d200** | 0.336979 | 0.461626 | 0.360657 | -.259116 | -.153957 | 0.656137 | -.112640 | -.096054 |
| **d400** | 0.355645 | 0.248273 | -.560467 | 0.652341 | -.218323 | 0.156625 | -.002854 | -.000127 |
| **d800** | 0.368684 | 0.012430 | -.532482 | -.479999 | 0.540053 | -.014692 | -.238016 | -.038165 |
| **d1500** | 0.372810 | -.139797 | -.153443 | -.404510 | -.487715 | -.157843 | 0.610011 | 0.139291 |
| **d5000** | 0.364374 | -.312030 | 0.189764 | 0.029588 | -.253979 | -.141299 | -.591299 | 0.546697 |
| **d10000** | 0.366773 | -.306860 | 0.181752 | 0.080069 | -.133176 | -.219017 | -.176871 | -.796795 |
| **marathon** | 0.341926 | -.438963 | 0.263209 | 0.299512 | 0.497928 | 0.315285 | 0.398822 | 0.158164 |

Y1 = 0.317556\* d100 + 0.336979\* d200 + 0.355645\* d400 + 0.368684\* d800 + 0.37281 \* d1500 + 0.364374\* d5000 + 0.366773\* d10000 + 0.341926\* marathon

Y2 = 0.566878 \* d100 + 0.461626 \* d200 + 0.248273 \* d400 + 0.01243 \* d800 -0.139797 \* d1500 - 0.31203 \* d5000 -0.30686 \* d10000 -0.438963 \* marathon

Interpretation:

Y1 - This component has more or less equal contribution from all of the variables. One reason here is that we have standardized the variables and therby reduced the impact of higher variance in longer races.

Y2 – This component has positive contributions from the smaller races (100m to 800m) and negative contributions from the longer races (1500 M to Marathons)

1. The highest and lowest scores for Principal component 1 are:

|  |  |  |
| --- | --- | --- |
| country | Prin1 correlation | Prin1 correlation with translated variable \* |
| usa | 86.07317 | -5.256791087 |
| cookis | 108.2157 | 16.88577235 |

\*uses the difference between the variables and their sample means

The highest and lowest scores for Principal component 2 are:

|  |  |  |
| --- | --- | --- |
| country | Prin2 correlation | Prin2 correlation with translated variable \* |
| wsamoa | -58.9458 | -12.39427182 |
| portugal | -42.119 | 4.432569597 |

\*uses the difference between the variables and their sample means

1. Comparison



Lets look at the following reasons one at a time:

* The results of principal component analysis depend on the scales at which the variables are measured.

>> Here the scales appear to be similar

* Variables with the highest sample variances will tend to be emphasized in the first few principal components when using covariance

>> In comparing races there may be some relevance of the higher variance in the longer races. If this variance is important and should be a part of the analysis, we should use the covariance matrix.

* Principal component analysis using the covariance function should only be considered if all of the variables have the same units of measurement.

>> All the variables are measured at the same scale (minutes) therefore covariance function can be considered.