Principal Component Analysis

gdito

4/9/2020

Note: output dari R pada dokumen ini diawali dengan tanda ##

Package

Pada Praktikum kali ini package yang dibutuhkan adalah

- factoextra
- ggcorrplot
- openxlsx

Silahkan install jika belum ada

```
install.packages("factoextra")
install.packages("ggcorrplot")
install.packages("openxlsx")

library(factoextra)
library(ggcorrplot)
```

Data Pelari Wanita

Berikut adalah data catatan waktu hasil tujuh nomor cabang lari atletik wanita yang berasal dari 55 negara pada salah satu event olimpiade yaitu lari 100 meter, 200 meter, 400 meter, 800 meter, 1500 meter, 3000 meter, dan maraton. Tiga nomor cabang lari pertama dicatat dalam satuan detik, sedangkan empat nomor yang lain dalam menit.

Tahap 1 Menyiapkan data di R

```
data_women_records <- openxlsx::read.xlsx("E:/APG/R
APG/women_track_records.xlsx")
head(data_women_records)

##    l00m  200m  400m 800m 1500m 3000m Marathon   country
## 1 11.61 22.94 54.50 2.15  4.43  9.79   178.52 argentina
## 2 11.20 22.35 51.08 1.98  4.13  9.08   152.37 australia
## 3 11.43 23.09 50.62 1.99  4.22  9.34   159.37   austria
## 4 11.41 23.04 52.00 2.00  4.14  8.88   157.85   belgium
## 5 11.46 23.05 53.30 2.16  4.58  9.81   169.98   bermuda
## 6 11.31 23.17 52.80 2.10  4.49  9.77   168.75    brazil
```

Note : `openxlsx::read.xlsx` berarti menggunakan fungsi `read.xlsx` yang berada pada package `openxlsx` tanpa memanggil packagenya terlebih dahulu menggunakan `library`. Fungsi `head` digunakan untuk menampilkan data 6 baris pertama.

Untuk keperluan analisis selanjutnya nama negara (country) akan dijadikan nama baris pada data. Hal ini dilakukan dengan menggunakan fungsi `rownames`.

```
rownames(data_women_records) <- data_women_records$country
data_women_records <- data_women_records[,-8]
```

`data_women_records[,-8]` berarti kita menghilangkan kolom kedelapan pada data (kolom country).

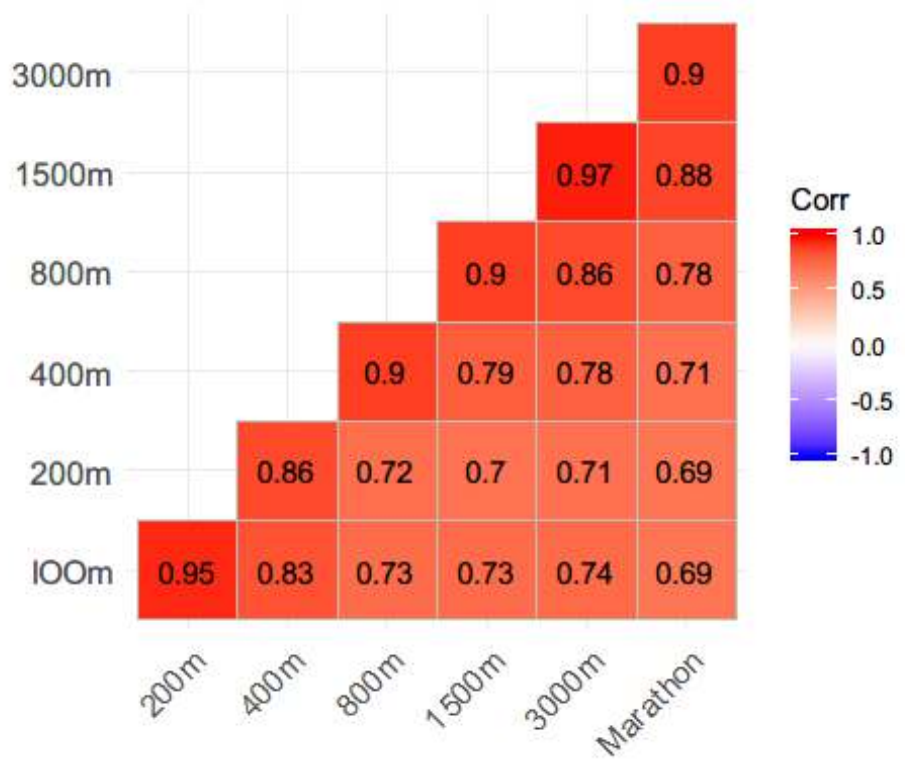
Tahap 2 Eksplorasi dengan menggunakan matrix korelasi

```
cor_women <- cor(data_women_records)
cor_women
```

```
##               l0Om      200m      400m      800m     1500m     3000m
## l0Om     1.0000000 0.9527911 0.8346918 0.7276888 0.7283709 0.7416988
## 200m     0.9527911 1.0000000 0.8569621 0.7240597 0.6983643 0.7098710
## 400m     0.8346918 0.8569621 1.0000000 0.8984052 0.7878417 0.7776369
## 800m     0.7276888 0.7240597 0.8984052 1.0000000 0.9016138 0.8635652
## 1500m    0.7283709 0.6983643 0.7878417 0.9016138 1.0000000 0.9691690
## 3000m    0.7416988 0.7098710 0.7776369 0.8635652 0.9691690 1.0000000
## Marathon 0.6863358 0.6855745 0.7054241 0.7792922 0.8779334 0.8998374
##           Marathon
## l0Om     0.6863358
## 200m     0.6855745
## 400m     0.7054241
## 800m     0.7792922
## 1500m    0.8779334
## 3000m    0.8998374
## Marathon 1.0000000
```

Agar mudah dilihat matrix korelasi ini bisa dibuat dalam bentuk grafik dengan cara berikut.

```
ggcorrplot(cor_women,type="lower",lab = TRUE)
```



Tahap 3 Menerapkan PCA (AKU)

Dalam R, Penerapan PCA ini dapat dilakukan dengan menggunakan `fungsi prcomp`. Fungsi ini memiliki argumen `scale.` dan `center`. Jika kedua argumen ini `TRUE` maka matrix yang digunakan untuk menghitung PCA adalah matrix korelasi. Namun, jika kedua argumen ini `FALSE` atau `scale.=FALSE`, maka matrix yang digunakan adalah matrix covariance.

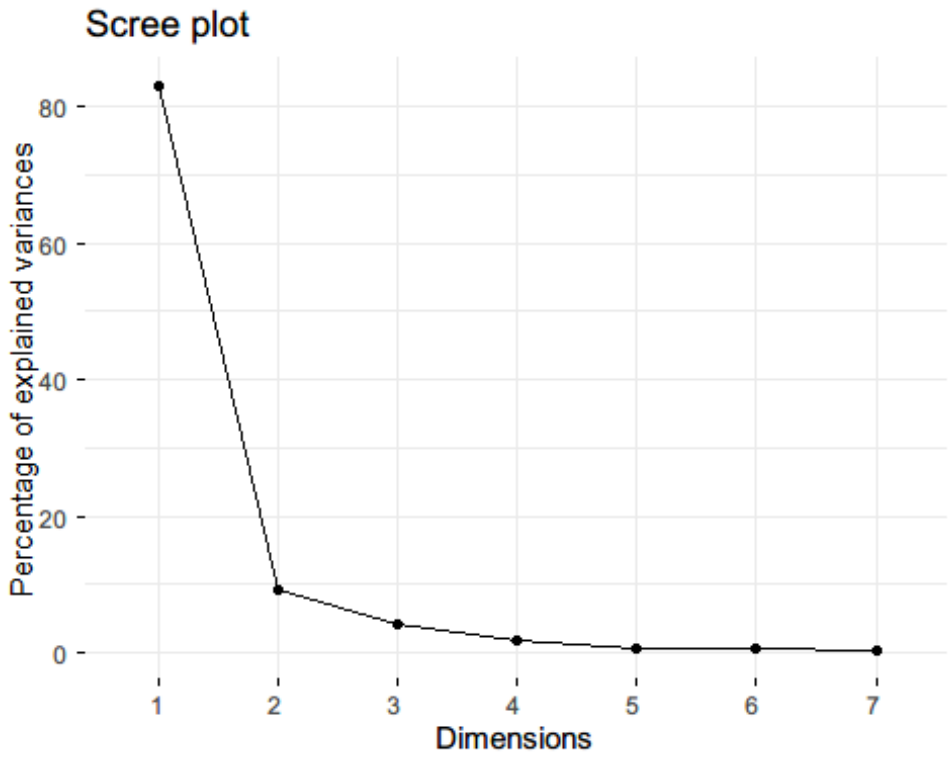
```
pca_women_records <- prcomp(data_women_records,scale.=TRUE,center=TRUE)
summary(pca_women_records)

## Importance of components:
##                           PC1     PC2     PC3     PC4     PC5     PC6
## Standard deviation     2.4095 0.80848 0.54762 0.35423 0.23198 0.19761
## Proportion of Variance 0.8294 0.09338 0.04284 0.01793 0.00769 0.00558
## Cumulative Proportion  0.8294 0.92276 0.96560 0.98353 0.99122 0.99679
##                           PC7
## Standard deviation     0.14981
## Proportion of Variance 0.00321
## Cumulative Proportion  1.00000
```

Hasil yang dikeluar dari sintaks diatas ada tiga macam yaitu Standard deviation, Proportion of Variance dan Cumulative Proportion dari masing-masing Komponen Utama (Principal Component). Standard deviaton merupakan akar dari akar ciri (eigenvalue). Dalam hal ini akar ciri berperan sebagai variance dari masing-masing komponen utama. Proportion of Variance didapatkan dari akar ciri pada masing-masing komponen dibagi dengan total akar ciri. Proportion of Variance menjelaskan seberapa besar keragaman peubah asal yang

dapat dijelaskan oleh masing-masing komponen utama. Semakin besar nilainya berarti semakin baik pula komponen utama tersebut untuk merepresentasikan peubah asal.

Cumulative Proportion menjelaskan seberapa besar keragaman yang dapat dijelaskan oleh komponen utama secara kumulatif. Misalnya saja dengan menggunakan dua komponen utama saja (PC1 dan PC2), sudah bisa menjelaskan lebih dari 92% keragaman dari data. Berdasarkan hal ini, kita akan memilih menggunakan dua komponen utama saja.

```
fviz_screeplot(pca_women_records,geom="line")
```



Hal lain yang bisa dilakukan untuk menentukan berapa banyak komponen utama yang digunakan adalah dengan screeplot. Fungsi untuk menampilkan screeplot pada R adalah `fviz_screeplot` yang didapat dari package `factoextra`.

Banyaknya komponen utama bisa ditentukan dengan screeplot dengan melihat di komponen utama yang mana garisnya berbentuk seperti siku (elbow). Pada gambar diatas garis membentuk siku saat berada di komponen utama kedua (dimension kedua). Sehingga banyaknya komponen utama yang digunakan sebanyak dua (Komponen Utama 1 dan Komponen Utama 2)

Tahap 4 Interpretasi PCA (AKU)

Interpretasi metode PCA dapat dilakukan dengan menggunakan vektor ciri pada masing-masing komponen utama. Semakin besar vektor ciri pada komponen utama tertentu maka semakin besar pula kontribusi dari peubah asal untuk membangun komponen utama tersebut. Catatan lain yang perlu diperhatikan adalah nilai negatif pada vektor ciri menandakan peubah asal memberikan kontribusi yang berkembalikan pada pembentukan

komponen utama. Dalam konteks vektor ciri negatif, semakin besar nilai peubah asal semakin kecil nilai pada komponen utama.

```
pca_women_records$rotation

##                PC1        PC2         PC3         PC4         PC5
## lOOm     0.3683561  0.4900597 -0.28601157  0.31938631  0.23116950
## 200m     0.3653642  0.5365800 -0.22981913 -0.08330196  0.04145457
## 400m     0.3816103  0.2465377  0.51536655 -0.34737748 -0.57217791
## 800m     0.3845592 -0.1554023  0.58452608 -0.04207636  0.62032379
## 1500m    0.3891040 -0.3604093  0.01291198  0.42953873  0.03026144
## 3000m    0.3888661 -0.3475394 -0.15272772  0.36311995 -0.46335476
## Marathon 0.3670038 -0.3692076 -0.48437037 -0.67249685  0.13053590
##                  PC6         PC7
## lOOm      0.619825234  0.05217655
## 200m     -0.710764580 -0.10922503
## 400m      0.190945970  0.20849691
## 800m     -0.019089032 -0.31520972
## 1500m    -0.231248381  0.69256151
## 3000m     0.009277159 -0.59835943
## Marathon  0.142280558  0.06959828
```

Karena kita hanya menggunakan dua komponen saja, maka vector ciri yang akan dinterpretasikan hanya pada PC1 dan PC2. PC1 memiliki vektor ciri yang relatif sama yaitu berkisar di 0.3 untuk semua cabang lomba. Vektor ciri yang relatif sama ini menandakan bahwa kontribusi peubah asal untuk membangun komponen utama ini relatif sama. Artinya nilai-nilai yang ada di PC1 (score value) dapat menggambarkan waktu lari untuk semua cabang lomba. Oleh karena itu kita dapat dapat menggunakan PC1 untuk menentukan negara mana yang memiliki pelari tercepat untuk semua kategori lomba. Vektor ciri di PC2 memiliki nilai positif untuk cabang lari jarak dekat (100m -400m) dan nilai negatif untuk cabang lari jarak jauh(800m-marathon). Hal ini berarti semakin besar score value pada PC2 maka waktu lari cabang jarak dekat semakin lambat namun waktu lari untuk cabang jarak jauh semakin cepat. Oleh karena itu, PC2 dapat digunakan untuk menentukan negara mana yang pada cabang lari jarak dekat waktunya mirip seperti cabang lari jarak jauh.

Note: Interpretasi komponen utama memiliki subjektifitas yang tinggi, oleh karena itu setiap orang menginterpretasikanya berbeda

Hal terakhir yang bisa diinterpretasikan adalah score value pada PC1 dan PC2. Score value merupakan observasi/koordinat baru pada peubah komponen utama. Dalam konteks data pelari diatas, observasinya adalah negara, sehingga kita dapat memberi insight cabang perlombaan lari dari setiap negara.

Untuk melihat score value pada komponen utama dapat dilihat dengan menggunakan sintaks berikut.

```
pca_women_records$x
```

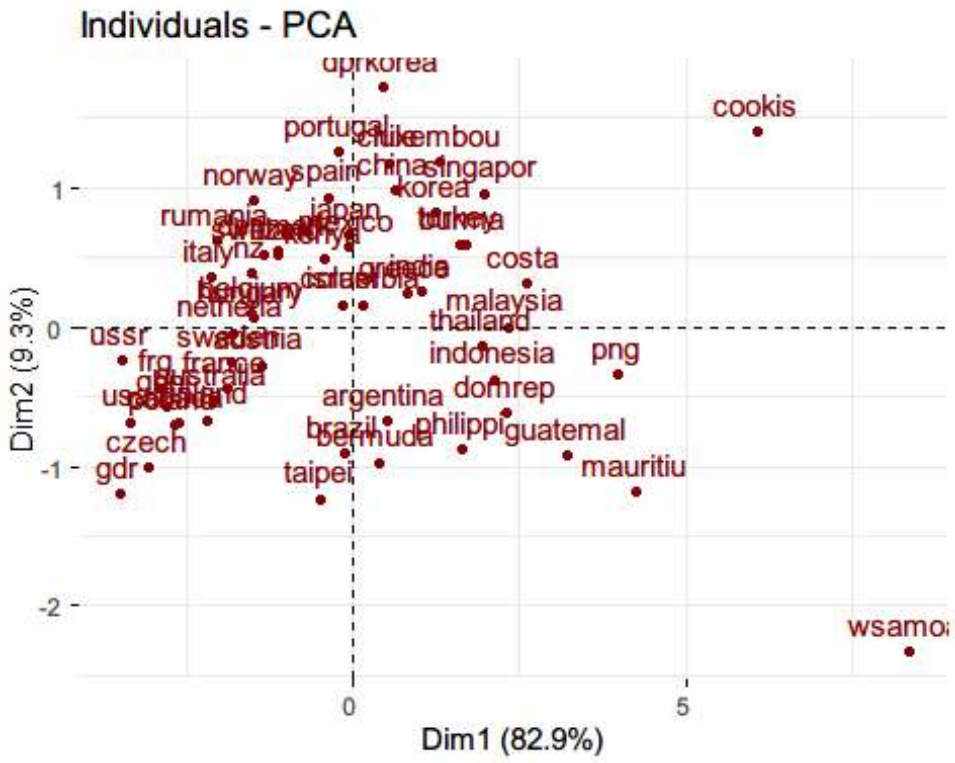
```
##                 PC1          PC2         PC3         PC4          PC5
## argentina  0.52726229 -0.674719057  0.61558926  0.04552931 -0.0025575859
## australia -2.09355119 -0.532798687 -0.04317487  0.18737792 -0.2183555124
## austria   -1.38044331 -0.274995495 -0.53231305  0.42623519 -0.0255039187
## belgium   -1.50998970  0.090963946 -0.08345684 -0.03942270 -0.0303263716
## bermuda    0.38781757 -0.976409482  0.64887210  0.47445652  0.2043216118
## brazil    -0.11839732 -0.911515514  0.32214131  0.34096557 -0.0959612766
## burma      1.68203844  0.586191540  0.07582590 -0.14511731  0.6034322854
## canada    -2.60813211 -0.695287897  0.10954223  0.03328484  0.1410820801
## chile      0.54783013  1.171926564 -0.23733316 -0.09612578 -0.2156317054
## china      0.64127320  0.980047440  0.05370738  0.02293887 -0.0579061799
## columbia   0.14157212  0.154468463  0.21456812  0.17614694  0.1895221521
## cookis     6.07727834  1.399503624 -0.21282983 -0.28085726 -0.0529187535
## costa      2.61922856  0.305956420  1.09460087  0.41203614 -0.5847203364
## czech     -3.05379905 -1.012730429 -1.04879114  0.37317120 -0.0258660649
## denmark   -1.11637538  0.540131452  0.41098527 -0.17591883 -0.1041309755
## domrep     2.29543640 -0.616097039  0.64633093 -0.26243482  0.3964092177
## finland   -2.18183995 -0.670248707  0.08087437  0.08706670  0.3299422783
## france    -1.89216992 -0.443832046  0.14465457 -0.05322917 -0.1896245118
## gdr       -3.50601681 -1.202500275 -0.52603497 -0.12430576  0.0884050075
## frg       -2.92577741 -0.437137537 -0.20120561 -0.02584472  0.0480344770
## gbni      -2.78315609 -0.578349737  0.13304952 -0.13024804  0.0417404253
## greece     0.81424542  0.233714829 -0.15991295 -0.08693903 -0.4548804436
## guatemal   3.22729824 -0.919057694  0.44304609 -0.09073846  0.3545878042
## hungary   -1.47721337  0.059384256 -0.15006131  0.12504871  0.0924438246
## india      1.01453673  0.253605564 -0.51071114  0.05275360  0.0509072296
## indonesia  2.11236466 -0.378269923  0.33669289 -0.18769360  0.3751519942
## ireland   -1.11735099  0.513043238  0.44713600 -0.08797261 -0.0255638127
## israel    -0.14296749  0.155867013  0.75136633 -0.16605776 -0.2905821673
## italy     -2.13954076  0.351991902 -0.07731676 -0.29052538 -0.2244816386
## japan     -0.05923268  0.657973909  0.40042678  0.42998430  0.1230753651
## kenya     -0.43089430  0.480611126 -0.75309705 -0.32602370 -0.0643810184
## korea      1.23386149  0.814398564  0.55640268  0.23959883  0.0129823046
## dprkorea   0.46229683  1.717885467 -1.91963681  0.34183702  0.3375496932
## luxembou   1.30174495  1.182174513 -0.10512035 -0.03151618 -0.4495547106
## malaysia   2.34053535 -0.001259817  0.11819610  0.84655411  0.1277242627
## mauritiu   4.23384717 -1.180202966 -0.08772947 -1.39411471 -0.1640608518
## mexico    -0.06348187  0.568969717 -0.09040891  0.45214535 -0.2961853317
## netherla  -1.79442661 -0.047085355  0.14270838 -0.10800877 -0.3625892311
## nz        -1.51125893  0.377920885  0.06112304  0.36901628  0.2627172276
## norway    -1.48300990  0.904195203  0.38741591 -0.14529920  0.1311735490
## png        3.98086034 -0.340244237 -0.28834865 -0.29398974  0.1595586421
## philippi   1.64018955 -0.876042797  0.22214354 -0.02227534  0.2056514197
## poland    -2.67209659 -0.702323548 -0.59597055 -0.02384024  0.0364171238
## portugal  -0.22428415  1.252662130  0.46217831 -0.02349218  0.2384703479
## rumania   -2.02982623  0.618046803 -0.83844940 -0.46958197 -0.0740123635
## singapor   1.97013151  0.951982007 -0.38929662  0.40329233  0.0732785524
## spain     -0.35565287  0.925478452 -0.05010366 -0.10341725  0.0927143013
## sweden    -1.82775494 -0.254820864  0.24805824 -0.14902201 -0.0006283576
## switzerl  -1.34665382  0.514180988  0.24518048 -0.22415740 -0.0858444392
```

```
## taipei    -0.50011940 -1.234653297  0.31159932 -0.04781148  0.0094018502
## thailand   1.95317730 -0.139873925  0.81360772  0.65551039 -0.1576314482
## turkey     1.60820411  0.594342560  0.16105018 -0.81523551  0.1477445097
## usa       -3.33581190 -0.685104574  0.38123893 -0.27057989 -0.1954822627
## ussr      -3.46468721 -0.245078447 -0.64637481 -0.20743799 -0.0824770130
## wsamoa     8.33288156 -2.326979228 -1.49263486  0.40428466 -0.3425812545
##                   PC6          PC7
## argentina  0.457907931 -0.079962873
## australia  0.137966927 -0.009815157
## austria   -0.081682704 -0.106172559
## belgium    0.062877332  0.138490873
## bermuda   -0.049426348  0.047829472
## brazil    -0.300448073 -0.006724259
## burma      0.277326099  0.055825566
## canada    -0.116436013 -0.117243747
## chile      0.128882585  0.003967955
## china      0.046618764  0.172340012
## columbia  -0.324561612 -0.122442477
## cookis    -0.055706058 -0.289275038
## costa     -0.065656582 -0.044687998
## czech      0.047409019  0.188230868
## denmark   -0.179964399  0.322413538
## domrep    -0.006777189  0.321472376
## finland   -0.031636754 -0.182734885
## france    -0.035807689  0.053223788
## gdr       -0.047152707 -0.176067724
## frg       -0.191479813  0.086729269
## gbni       0.012368224  0.059974475
## greece     0.093873143 -0.121106747
## guatemal  -0.279702884 -0.030512752
## hungary    0.061704372 -0.002880404
## india      0.134490600 -0.442230795
## indonesia -0.013434970 -0.058448424
## ireland   -0.149414709 -0.043637389
## israel    -0.090944777 -0.094991490
## italy      0.012049601  0.080196477
## japan     -0.182028982  0.141996717
## kenya      0.119151878 -0.011054538
## korea     -0.028182131 -0.027699883
## dprkorea  -0.561883421 -0.072075631
## luxembou  -0.115327186  0.123404789
## malaysia   0.374021886 -0.140160623
## mauritiu  -0.321676670 -0.209056067
## mexico     0.402826018 -0.127280417
## netherla   0.051567793 -0.071966309
## nz         0.078135023  0.198553009
## norway     0.226377016  0.086297791
## png        0.111686449  0.039358192
## philippi   0.267820061 -0.094783185
## poland     0.144920475 -0.248593496
```

```
## portugal  -0.041084319  0.041047319
## rumania   -0.050795933  0.167824160
## singapor   0.091437114  0.018007615
## spain      0.124420761 -0.021530131
## sweden    -0.159842868  0.036333974
## switzerl   0.047330468  0.068598802
## taipei     0.024237079 -0.093082005
## thailand  -0.482014923 -0.032198783
## turkey     0.287104548  0.191270368
## usa       -0.047049486  0.040368337
## ussr       0.097960158  0.017756855
## wsamoa     0.087647875  0.376903191
```

Agar lebih mudah dalam menginterpretasikan score value maka digunakaan grafik dibawah ini.

```
fviz_pca_ind(pca_women_records,col.ind = "darkred")
```



Berdasarkan grafik score value dapat diketahui bahwa negara yang memiliki catatan waktu pelari terlambat untuk semua cabang lomba adalah negara wsamoa. Hal ini dikarenakan wsamoa score value wsakoa untuk PC1 (Dim1) paling besar diantara yang lain. Walaupun negara wsamoa memiliki cabang lari terlama disemua cabang lomba, namun perbedaan waktu terkecil antara pelari jarak jauh dan jarak dekat adalah negara wsamoa. Hal ini berarti pelari untuk lomba jarak dekat sangat lambat karena memiliki waktu yang hampir mirip seperti pelari jarak jauh. Sedangkan negara yang memiliki pelari tercepat untuk semua cabang lomba adalah gdr.

Demikianlah contoh interpretasi dari PCA, apakah anda punya interpretasi lain?