Factor_Analysis

Sachith M Gunawardane

2023-06-11

#Load data #Data file is available under the same Git folder

```
data <- read.table("D:/PGIS_Data_Science_AI/DS5110_Statistical_simulation/projects/Factor_Analysis/trac
head(data)
```

```
##         C1.T X100m X200m X400m X800m X1500m X5000m X10000m Marathon
## 1 Argentina 10.23 20.37 46.18  1.77   3.68  13.33   27.65   129.57
## 2 Australia  9.93 20.06 44.38  1.74   3.53  12.93   27.53   127.51
## 3   Austria 10.15 20.45 45.80  1.77   3.58  13.26   27.72   132.22
## 4   Belgium 10.14 20.19 45.02  1.73   3.57  12.83   26.87   127.20
## 5   Bermuda 10.27 20.30 45.26  1.79   3.70  14.64   30.49   146.37
## 6    Brazil 10.00 19.89 44.29  1.70   3.57  13.48   28.13   126.05
```

```
dim(data)
```

```
## [1] 54  9
```

```
dataset1 <- data[,-1]
head(dataset1)
```

```
##   X100m X200m X400m X800m X1500m X5000m X10000m Marathon
## 1 10.23 20.37 46.18  1.77   3.68  13.33   27.65   129.57
## 2  9.93 20.06 44.38  1.74   3.53  12.93   27.53   127.51
## 3 10.15 20.45 45.80  1.77   3.58  13.26   27.72   132.22
## 4 10.14 20.19 45.02  1.73   3.57  12.83   26.87   127.20
## 5 10.27 20.30 45.26  1.79   3.70  14.64   30.49   146.37
## 6 10.00 19.89 44.29  1.70   3.57  13.48   28.13   126.05
```

#Check for Correlation

```
cor(dataset1)
```

```
##              X100m     X200m     X400m     X800m    X1500m    X5000m   X10000m
## X100m    1.0000000 0.9147554 0.8041147 0.7119388 0.7657919 0.7398803 0.7147921
## X200m    0.9147554 1.0000000 0.8449159 0.7969162 0.7950871 0.7613028 0.7479519
## X400m    0.8041147 0.8449159 1.0000000 0.7677488 0.7715522 0.7796929 0.7657481
## X800m    0.7119388 0.7969162 0.7677488 1.0000000 0.8957609 0.8606959 0.8431074
## X1500m   0.7657919 0.7950871 0.7715522 0.8957609 1.0000000 0.9165224 0.9013380
## X5000m   0.7398803 0.7613028 0.7796929 0.8606959 0.9165224 1.0000000 0.9882324
## X10000m  0.7147921 0.7479519 0.7657481 0.8431074 0.9013380 0.9882324 1.0000000
## Marathon 0.6764873 0.7211157 0.7126823 0.8069657 0.8777788 0.9441466 0.9541630
##           Marathon
## X100m    0.6764873
## X200m    0.7211157
## X400m    0.7126823
## X800m    0.8069657
## X1500m   0.8777788
## X5000m   0.9441466
## X10000m  0.9541630
## Marathon 1.0000000
```

As you can see data is highly /significantly correlated

#Bartlett.test This is to verify is there any possibility to do factor analysis

```
bartlett.test(dataset1)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  dataset1
## Bartlett's K-squared = 1435.7, df = 7, p-value < 2.2e-16
```

p-value is small ~ 0; which mean significant
hence factor analysis is possible

#Check all variables are good for factor analysis or not

```
library(psych)
```

```
## Warning: package 'psych' was built under R version 4.2.3
```

```
KMO(cor(dataset1))
```

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = cor(dataset1))
## Overall MSA =  0.89
## MSA for each item =
##    X100m    X200m    X400m    X800m   X1500m   X5000m  X10000m Marathon
##     0.84     0.84     0.97     0.90     0.94     0.85     0.85     0.95
```

1st - Overall MSA (Measure of Sampling Adequacy) = 0.89 If MSA is < 0.5, it indicate that overall factor analysis is not possible

2nd - Individual MSA also above 0.5 hence all variables are good for Factor analysis

#FACTOR analysis

```
factor1.out <-  factanal(dataset1, factors = 1 )
factor1.out
```

```
##
## Call:
## factanal(x = dataset1, factors = 1)
##
## Uniquenesses:
##    X100m    X200m    X400m    X800m   X1500m   X5000m  X10000m Marathon
##    0.446    0.404    0.383    0.251    0.152    0.009    0.017    0.094
##
## Loadings:
##          Factor1
## X100m    0.744
## X200m    0.772
## X400m    0.786
## X800m    0.865
## X1500m   0.921
## X5000m   0.996
## X10000m  0.992
## Marathon 0.952
##
##                Factor1
## SS loadings      6.245
## Proportion Var   0.781
##
## Test of the hypothesis that 1 factor is sufficient.
## The chi square statistic is 118.31 on 20 degrees of freedom.
## The p-value is 5.85e-16
```

It is important to look at P-Value H0: is One Factor is sufficient P-Value < 0.05 hence we reject H0 and conclude that 1 factor is not sufficient for this data set

```
factor2.out <-  factanal(dataset1, factors = 2 )
factor2.out
```

```
##
## Call:
## factanal(x = dataset1, factors = 2)
##
## Uniquenesses:
##    X100m    X200m    X400m    X800m   X1500m   X5000m  X10000m Marathon
##    0.135    0.037    0.228    0.212    0.134    0.012    0.011    0.088
##
## Loadings:
##          Factor1 Factor2
## X100m    0.397   0.841
## X200m    0.404   0.894
## X400m    0.511   0.714
## X800m    0.667   0.585
## X1500m   0.745   0.558
## X5000m   0.883   0.455
## X10000m  0.897   0.429
## Marathon 0.863   0.410
##
##                Factor1 Factor2
## SS loadings      3.912   3.231
## Proportion Var   0.489   0.404
## Cumulative Var   0.489   0.893
##
## Test of the hypothesis that 2 factors are sufficient.
## The chi square statistic is 25.94 on 13 degrees of freedom.
## The p-value is 0.0173
```

P-value < 0.05 therefore we will reject H0 Hence 2 factor solution is not sufficient for this

```
factor3.out <-  factanal(dataset1, factors = 3 )
factor3.out
```

```
##
## Call:
## factanal(x = dataset1, factors = 3)
##
## Uniquenesses:
##    X100m    X200m    X400m    X800m   X1500m   X5000m  X10000m Marathon
##    0.082    0.069    0.229    0.005    0.110    0.015    0.006    0.086
##
## Loadings:
##          Factor1 Factor2 Factor3
## X100m    0.366   0.866   0.187
## X200m    0.374   0.829   0.322
## X400m    0.472   0.676   0.302
## X800m    0.538   0.441   0.715
## X1500m   0.671   0.494   0.443
## X5000m   0.842   0.426   0.307
## X10000m  0.870   0.400   0.278
## Marathon 0.837   0.377   0.266
##
##                Factor1 Factor2 Factor3
## SS loadings      3.403   2.816   1.179
## Proportion Var   0.425   0.352   0.147
## Cumulative Var   0.425   0.777   0.925
##
## Test of the hypothesis that 3 factors are sufficient.
## The chi square statistic is 9.44 on 7 degrees of freedom.
## The p-value is 0.223
```

P-value > 0.05 therefore we are fail to reject H0 Hence 3 factor solution is sufficient for this

factor1 : X1500m & X5000m & X10000m and Marathon. Indicate all for longer distance runs factor2: 100m, 200m, 400m which are short distance events factor3: 800m - mid distance

```
factor3.out <-  factanal(dataset1, factors = 2, rotation = "none" )
factor3.out
```

```
##
## Call:
## factanal(x = dataset1, factors = 2, rotation = "none")
##
## Uniquenesses:
```

```
##    X100m    X200m    X400m    X800m   X1500m   X5000m  X10000m Marathon
##    0.135    0.037    0.228    0.212    0.134    0.012    0.011    0.088
##
## Loadings:
##          Factor1 Factor2
## X100m     0.780   0.507
## X200m     0.814   0.548
## X400m     0.811   0.338
## X800m     0.875   0.146
## X1500m    0.927
## X5000m    0.991
## X10000m   0.989  -0.107
## Marathon  0.949  -0.105
##
##                Factor1 Factor2
## SS loadings      6.415   0.728
## Proportion Var   0.802   0.091
## Cumulative Var   0.802   0.893
##
## Test of the hypothesis that 2 factors are sufficient.
## The chi square statistic is 25.94 on 13 degrees of freedom.
## The p-value is 0.0173
```

when we set rotation to varimax it try to optimize/maximize the variance this is useful when variables are hard/difficult to interpret

#How to identify # of factors required

```r
eigen(cor(dataset1))
```

```
## eigen() decomposition
## $values
## [1] 6.703289951 0.638410110 0.227524494 0.205849181 0.097577441 0.070687912
## [7] 0.046942050 0.009718862
##
## $vectors
##            [,1]        [,2]         [,3]        [,4]        [,5]        [,6]
## [1,] -0.3323877 -0.52939911 -0.343859303  0.38074525  0.29967117 -0.36203713
## [2,] -0.3460511 -0.47039050  0.003786104  0.21702322 -0.54143422  0.34859224
## [3,] -0.3391240 -0.34532929  0.067060507 -0.85129980  0.13298631  0.07708385
## [4,] -0.3530134  0.08945523  0.782711152  0.13427911 -0.22728254 -0.34130845
## [5,] -0.3659849  0.15365241  0.244270040  0.23302034  0.65162403  0.52977961
## [6,] -0.3698204  0.29475985 -0.182863147 -0.05462441  0.07181636 -0.35914382
## [7,] -0.3659489  0.33360619 -0.243980694 -0.08706927 -0.06133263 -0.27308617
## [8,] -0.3542779  0.38656085 -0.334632969  0.01812115 -0.33789097  0.37516986
##            [,7]         [,8]
## [1,]  0.3476470 -0.065701445
## [2,] -0.4398969  0.060755403
## [3,]  0.1135553 -0.003469726
## [4,]  0.2588830 -0.039274027
## [5,] -0.1470362 -0.039745509
## [6,] -0.3283202  0.705684585
## [7,] -0.3511133 -0.697181715
## [8,]  0.5941571  0.069316891
```

# of fators required should eigen values > 1

```r
library(FactoMineR)
```

```
## Warning: package 'FactoMineR' was built under R version 4.2.3
```

```r
library(factoextra)
```

```
## Warning: package 'factoextra' was built under R version 4.2.3
```

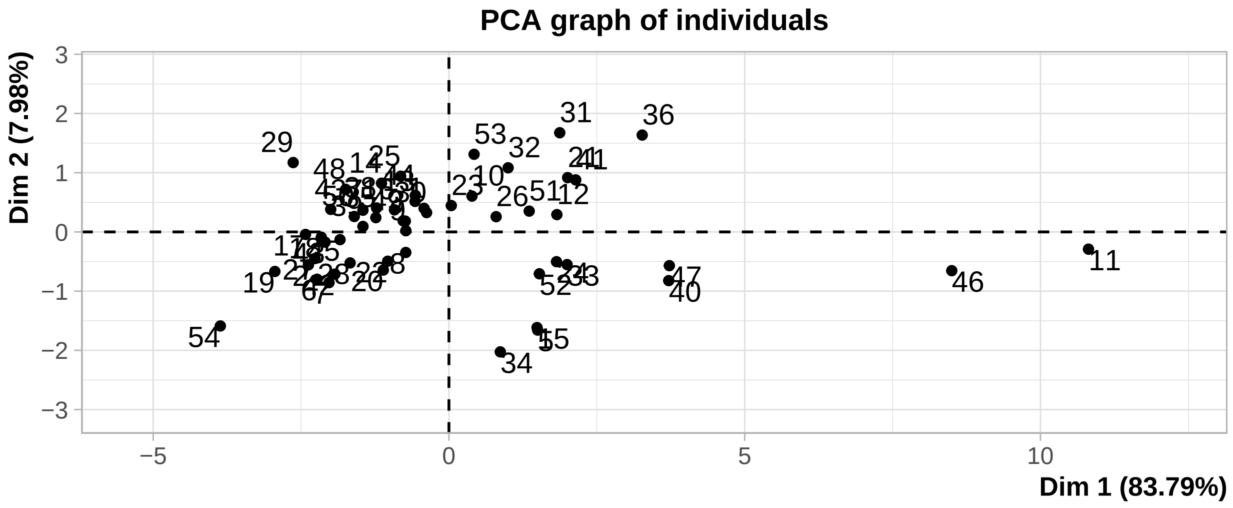
```
## Loading required package: ggplot2
```

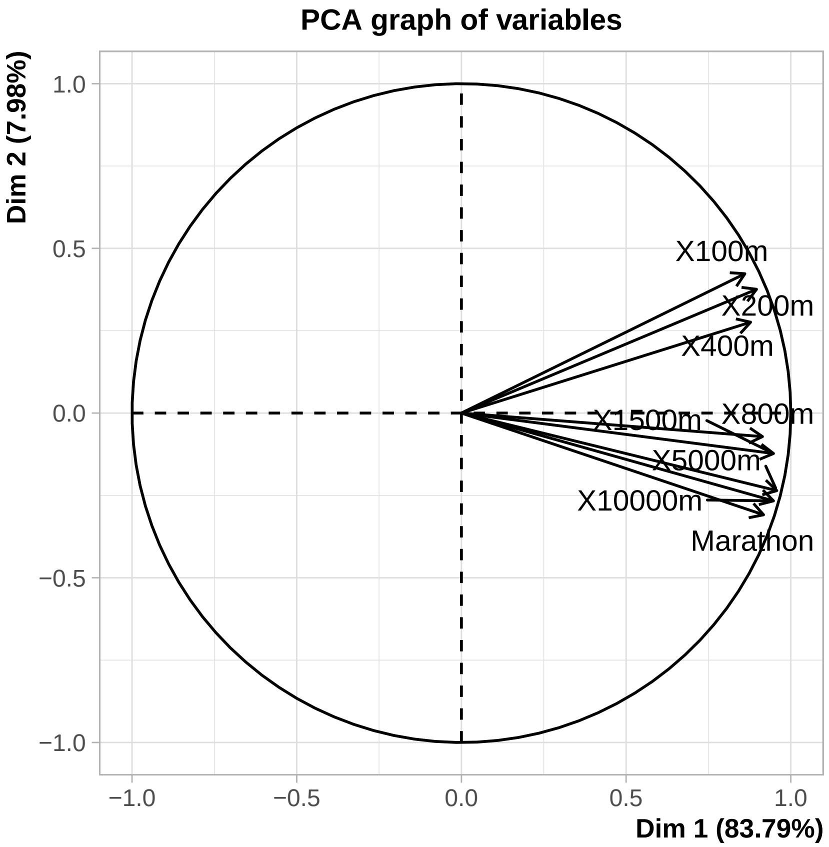
```
##
## Attaching package: 'ggplot2'
```

```
## The following objects are masked from 'package:psych':
##
##     %+%, alpha
```

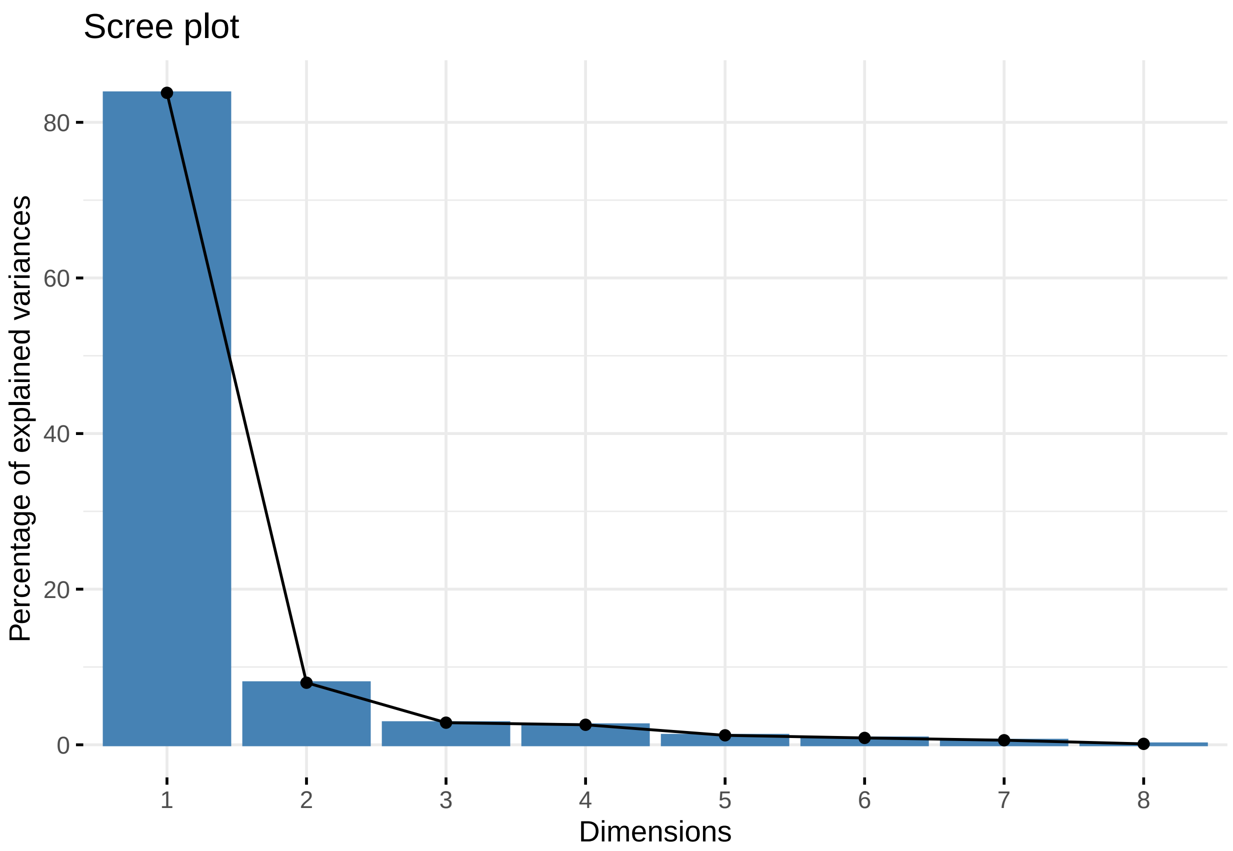
```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```r
pca.out <-  PCA(dataset1, ncp = 8)
```





```
fviz_eig(pca.out)
```



#Another Function for Factor Analysis

```
fa(cor(dataset1), nfactors = 1, rotate = "none", fm = "ml")
```

```
## Factor Analysis using method =  ml
## Call: fa(r = cor(dataset1), nfactors = 1, rotate = "none", fm = "ml")
## Standardized loadings (pattern matrix) based upon correlation matrix
##           ML1   h2     u2 com
## X100m    0.74 0.55 0.4460   1
## X200m    0.77 0.60 0.4039   1
## X400m    0.79 0.62 0.3828   1
## X800m    0.87 0.75 0.2512   1
## X1500m   0.92 0.85 0.1518   1
## X5000m   1.00 0.99 0.0088   1
## X10000m  0.99 0.98 0.0168   1
## Marathon 0.95 0.91 0.0938   1
##
##                 ML1
## SS loadings    6.24
## Proportion Var 0.78
##
## Mean item complexity =  1
## Test of the hypothesis that 1 factor is sufficient.
##
## df null model =  28  with the objective function =  14.28
## df of  the model are 20  and the objective function was  2.42
```

```
##
## The root mean square of the residuals (RMSR) is  0.1
## The df corrected root mean square of the residuals is  0.12
##
## Fit based upon off diagonal values = 0.99
## Measures of factor score adequacy
##                                                    ML1
## Correlation of (regression) scores with factors   1.00
## Multiple R square of scores with factors          0.99
## Minimum correlation of possible factor scores     0.99
```