Multivariate Analysis

Factor Analysis on Sport Events

GUIDED BY - DR. SANTOSHA C D

PREPARED BY - APOORVA GUPTA
NISTHA RULANIA
SUHASI GOHIL

Table of contents

Part 1:

Introduction

MVA & PCA

Multivariate Analysis

Multivariate means involving multiple dependent variables resulting in one outcome.

Example: To predict the sales of the company, we cannot simply say that 'cost' is the factor which will affect the sales. There are multiple variables which will impact sales. To analyze the variables that will impact sales majorly, can only be found with multivariate analysis.

Sales will depend on the category of product, production capacity, geographical location, marketing effort, competitor analysis, cost of the product etc.

Advantages:

It considers more than one factor of independent variables that influence the variability of dependent variables, the conclusion drawn is more accurate.

Ability to glean a more realistic picture than looking at a single variable.

Multivariate techniques provide a powerful test of significance compared to univariate techniques

Disadvantages:

Multivariate techniques are complex and involve high level mathematics that require a statistical program to analyze the data. These statistical programs can be expensive for an individual to obtain.

Statistical modeling outputs are not always easy for students to interpret.

For multivariate techniques to give meaningful results, they need a large sample of data; otherwise, the results are meaningless due to high standard errors.

MVA techniques

01 MULTIVARIATE LINEAR REGRESSION

02 LOGISTIC REGRESSION

03 FACTOR ANALYSIS

04 CLUSTER ANALYSIS

05 DISCRIMINANT ANALYSIS

06 MULTIVARIATE ANALYSIS OF VARIANCE (MANOVA)

Principal Component Analysis (PCA)

Factor Analysis is a useful approach to find latent variables which are not directly measured in a single variable but rather inferred from other variables in the dataset. These latent variables are called factors.

Principal Component Analysis, or PCA, is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set.

It was invented by Karl Pearson in 1901.

Assumptions in PCA

- Linearity in the data set
- Principal component with high variance must be paid attention and the PCs with lower variance are disregarded.
- All variables should be accessed on the same ratio level of measurement.
- Outliers or extreme values should be less.
- The feature set must be correlated and the reduced feature set after applying PCA will represent the original data set but in an effective way with fewer dimensions.

When & Why to use PCA ?

- PCA technique is particularly useful in processing data where multicollinearity exists between the features/variables.

- PCA can be also used for denoising and data compression.

- To also find the latent variables (which are not directly measured in a single variable but rather inferred from other variables in the dataset)

Applications of PCA

01 NEUROSCIENCE

02 QUANTITATIVE FINANCE

03 IMAGE COMPRESSION

04 FACIAL RECOGNITION

Project details

What is the secondary data about?

The data we use here is national track events data for men representing 54 countries in eight different events such as 100m, 800m, marathon etc.

Methodology

After proper cleaning of the dataset, we import the data and perform principal component analysis through R programming, Python and SPSS.

Data specifics

The data has 54 rows representing different countries like Argentina, India, Russia with 8 columns representing different races like 100m, 200m, 10,000m etc.

The dataset looks like as follows:

	A	B	C	D	E	F	G	H	I
1	**Country**	**X1**	**X2**	**X3**	**X4**	**X5**	**X6**	**X7**	**X8**
2	**Argentina**	10.23	20.37	46.18	1.77	3.68	13.33	27.65	129.57
3	**Australia**	9.93	20.06	44.38	1.74	3.53	12.93	27.53	127.51
4	**Austria**	10.15	20.45	45.8	1.77	3.58	13.26	27.72	132.22
5	**Belgium**	10.14	20.19	45.02	1.73	3.57	12.83	26.87	127.2
6	**Bermuda**	10.27	20.3	45.26	1.79	3.7	14.64	30.49	146.37
7	**Brazil**	10	19.89	44.29	1.7	3.57	13.48	28.13	126.05
8	**Canada**	9.84	20.17	44.72	1.75	3.53	13.23	27.6	130.09
9	**Chile**	10.1	20.15	45.92	1.76	3.65	13.39	28.09	132.19
10	**China**	10.17	20.42	45.25	1.77	3.61	13.42	28.17	129.18
11	**Columbia**	10.29	20.85	45.84	1.8	3.72	13.49	27.88	131.17

The description of the variables is as follows:

X1: 100m(s)

X2: 200m(s)

X3: 400m(s)

X4: 800m(min)

X5: 1500m(min)

X6: 5000m(min)

X7: 10000m(min)

X8: Marathon(min)

Part 2:

Objectives

MEN TRACK RECORDS

Objectives

Objective 1

To perform principal component analysis and **form clusters of countries** based on the factors extracted.

Objective 2

To also find the **latent variables for the different races** which represent the data well.

Objective 3

To analyze and compare the results obtained from **R programming, Python and SPSS.**

Part 3:

Hypotheses

THE NULL AND ALTERNATIVE HYPOTHESIS

Cofficient of correlation

H_0 : Correlation coefficients are not statistically significant

Vs

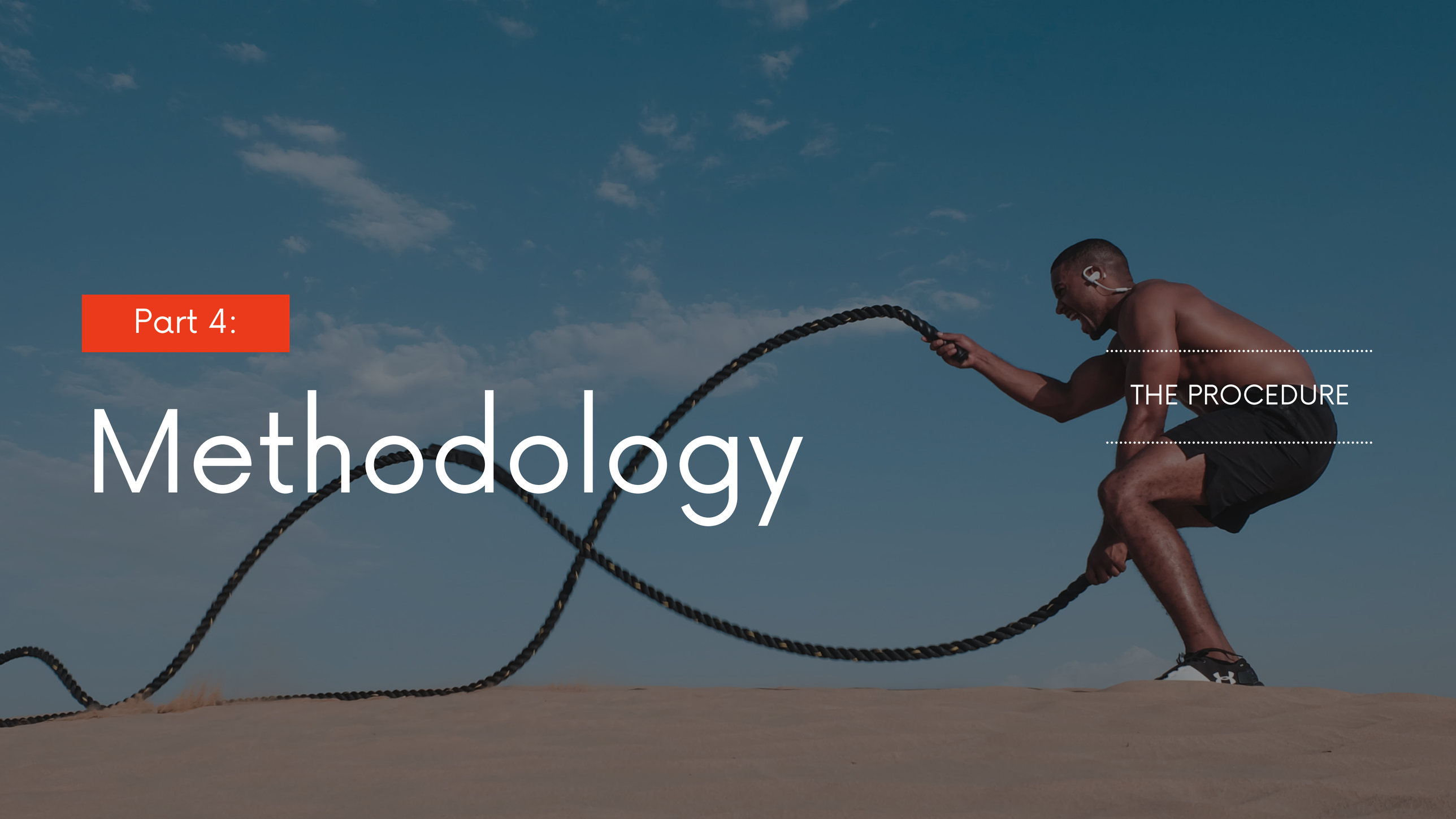
H_1 : Correlation coefficients are statistically significant

Sample adequacy

H_0 : The model is not statistically significant

Vs

H_1 : The model is statistically significant



Part 4:

Methodology

THE PROCEDURE

Select the best number of factors for the dataset

We use the correlation matrix to derive the factors. Factors derived from the correlation matrix are the same as those derived from the covariance matrix of the standardized (scaled) variables.

Eigen values determine the variance explained by each factor. We graph a scree plot which is a visual representation of the eigen values Vs the factors

According to Kaiser's rule, it is recommended to keep the factors with eigenvalues greater than 1.0. But it can also depend on the cumulative variance explained by the factors.

To find whether a factor rotation is needed for the data

Factor loadings indicate how much a factor explains a variable. They can range from -1 to 1.

Sometimes, the loadings on each factor do not spread very well and therefore they do not yield interpretable factors.

Hence, we use factor rotation which is an orthogonal transformation of original factors. This is done solely for the purpose of interpretation

Factor loadings, communalities and specific variances

Communalities determine the variability of each variable explained by the factors chosen.

Specific variance determines the effect of the factor on the specific variable and not shared with other variables.

We identify the variables on the basis of loadings in each factor. Using domain knowledge, relevant names can be given to the chosen factors.

Part 5:

Results

SOFTWARE CODES AND OUTPUT

R code for PCA

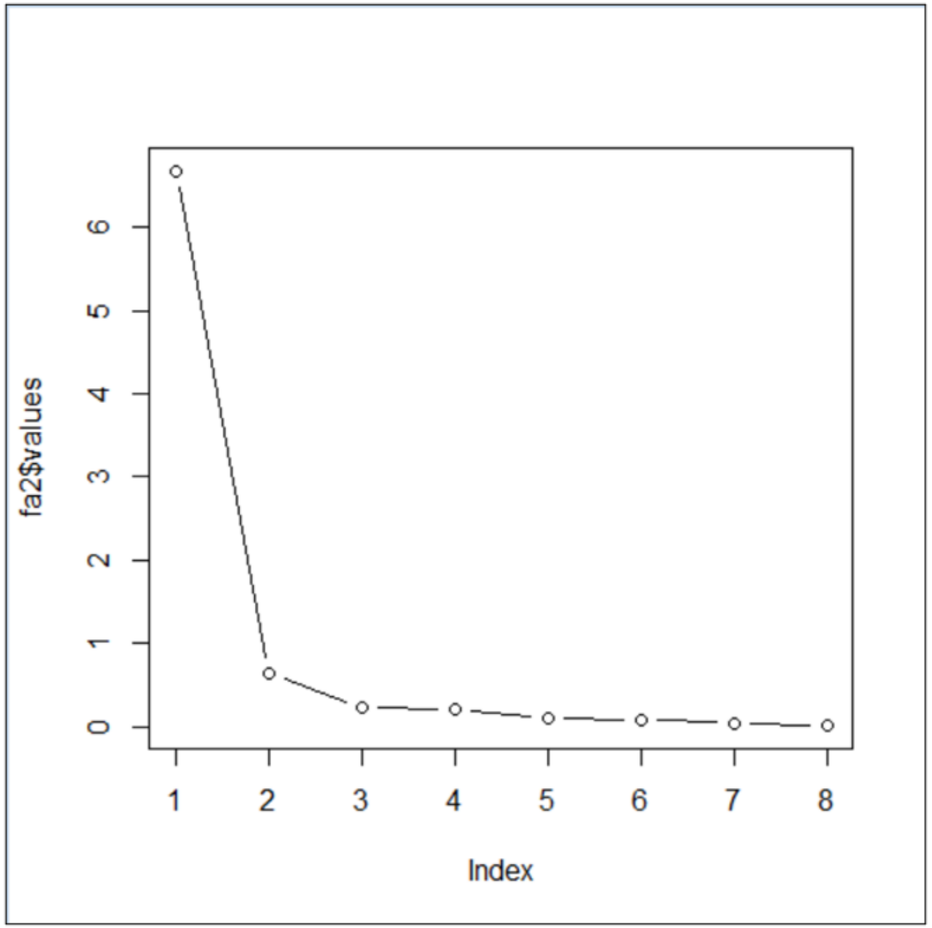
```
principal component analysis -

x <- read.csv("C:/Users/HP/OneDrive/Documents/mva data - men.csv")
x
y <- x[,2:9]
y


fa2 <- principal(y,nfactors = 4,rotate = "none",covar = FALSE)
fa2$values
plot(fa2$values, type = "b")
fa3 <- principal(y,nfactors = 2,rotate = "none",covar = FALSE)
fa3$loadings
fa4 <- principal(y,nfactors = 2,rotate = "varimax",covar = FALSE)
fa4$loadings
fa4
```

R PROGRAMMING OUTPUT

Eigen values:

```
> fa2 <- principal(y,nfactors = 4,rotate = "none",covar = FALSE)
> fa2$values
[1] 6.68112634 0.63747585 0.23168714 0.21295129 0.09557688 0.08123144 0.04771277
[8] 0.01223829
> 
```

Scree plot:



Original factor loadings

```
> fa3 <- principal(y,nfactors = 2,rotate = "none",covar = FALSE)
> fa3$loadings

Loadings:
    PC1    PC2
X1  0.862  0.418
X2  0.898  0.373
X3  0.876  0.278
X4  0.913
X5  0.941 -0.108
X6  0.957 -0.238
X7  0.943 -0.276
X8  0.917 -0.312

                 PC1   PC2
SS loadings    6.681 0.637
Proportion Var 0.835 0.080
Cumulative Var 0.835 0.915
> 
```

Rotated factor loadings

```
> fa4 <- principal(y,nfactors = 2,rotate = "varimax",covar = FALSE)
> fa4$loadings

Loadings:
    RC1   RC2
X1 0.370 0.883
X2 0.427 0.873
X3 0.474 0.788
X4 0.730 0.552
X5 0.778 0.541
X6 0.876 0.454
X7 0.889 0.416
X8 0.894 0.372

                 RC1   RC2
SS loadings    4.039 3.280
Proportion Var 0.505 0.410
Cumulative Var 0.505 0.915
> 
```

Factor loadings, communalities and specific variances

```
> fa4
Principal Components Analysis
Call: principal(r = y, nfactors = 2, rotate = "varimax", covar = FALSE)
Standardized loadings (pattern matrix) based upon correlation matrix
    RC1  RC2   h2    u2 com
X1 0.37 0.88 0.92 0.082 1.3
X2 0.43 0.87 0.95 0.055 1.5
X3 0.47 0.79 0.85 0.155 1.6
X4 0.73 0.55 0.84 0.162 1.9
X5 0.78 0.54 0.90 0.102 1.8
X6 0.88 0.45 0.97 0.028 1.5
X7 0.89 0.42 0.96 0.036 1.4
X8 0.89 0.37 0.94 0.062 1.3

                       RC1  RC2
SS loadings           4.04 3.28
Proportion Var        0.50 0.41
Cumulative Var        0.50 0.91
Proportion Explained  0.55 0.45
Cumulative Proportion 0.55 1.00
```

Python code for PCA (factor loadings) -

```
import pandas as pd
df = pd.read_csv("../input/men-data/men_data.csv")

X = df.iloc[:, 1:8]

from sklearn.preprocessing import StandardScaler
sc = StandardScaler()
X_scaled = sc.fit_transform(X)

from factor_analyzer import FactorAnalyzer
fa = FactorAnalyzer(n_factors=2, rotation="varimax", method="principal",
                    is_corr_matrix=False)
fa.fit(X_scaled)

print("Eigenvalues:")
print(fa.get_eigenvalues()[0])
print()
print("Communalities:")
print(fa.get_communalities())
print()
print("Specific Variances:")
print(fa.get_uniquenesses())
print()
print("Factor Loadings:")
print(fa.loadings_)
```

Python code for PCA (scree plot) -

```
import matplotlib.pyplot as plt
plt.style.use("ggplot")

plt.plot(fa.get_eigenvalues()[0], marker='o')
plt.xlabel("Eigenvalue number")
plt.ylabel("Eigenvalue size")
plt.title("Scree Plot")
```

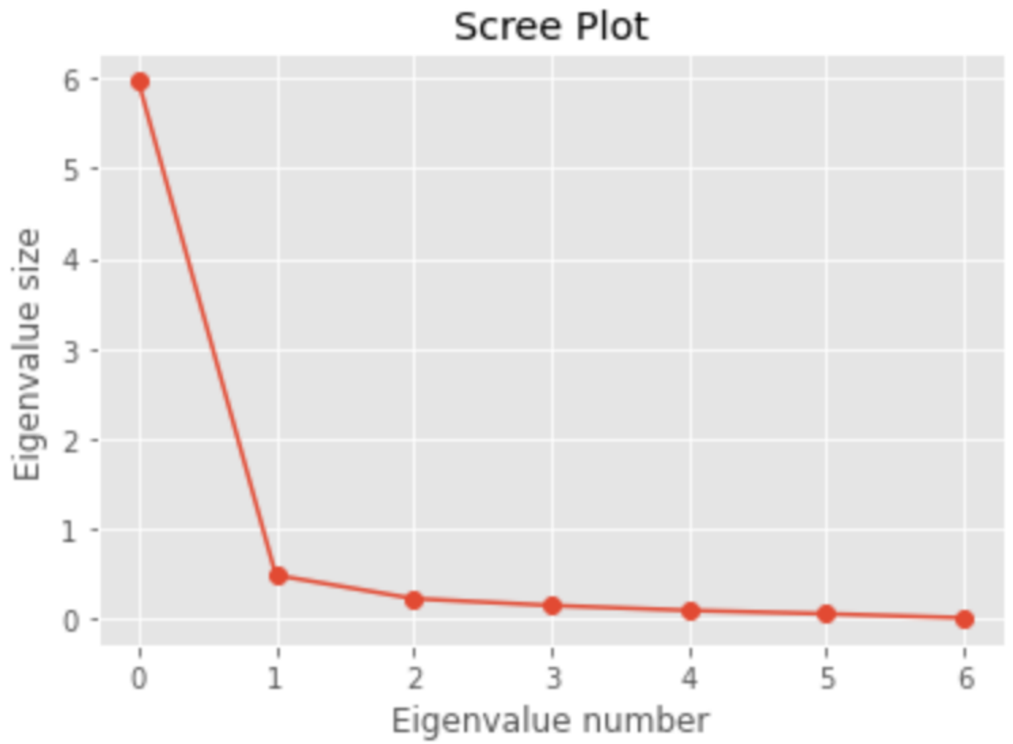
PYTHON OUTPUT

Model:

```
Eigenvalues:
[5.97153883 0.48638736 0.22631702 0.15051339 0.09530889 0.0574633
 0.01247121]

Communalities:
[0.92295827 0.89883129 0.85065557 0.89784503 0.97449894 0.96866515
 0.94447194]

Specific Variances:
[0.07704173 0.10116871 0.14934443 0.10215497 0.02550106 0.03133485
 0.05552806]

Factor Loadings:
[[0.41055459 0.86856387]
 [0.41318514 0.85329323]
 [0.67464557 0.628895  ]
 [0.75916011 0.56702818]
 [0.86678817 0.47241636]
 [0.88368065 0.43332858]
 [0.89078243 0.38855965]]
```

Scree plot:

```
[4]: Text(0.5, 1.0, 'Scree Plot')
```



SPSS OUTPUT

Correlation matrix:

		X1	X2	X3	X4	X5	X6	X7	X8
Correlation	X1	1.000	.915	.804	.712	.764	.740	.714	.676
	X2	.915	1.000	.845	.797	.806	.761	.739	.721
	X3	.804	.845	1.000	.768	.757	.780	.758	.713
	X4	.712	.797	.768	1.000	.885	.861	.835	.807
	X5	.764	.806	.757	.885	1.000	.904	.886	.867
	X6	.740	.761	.780	.861	.904	1.000	.986	.944
	X7	.714	.739	.758	.835	.886	.986	1.000	.951
	X8	.676	.721	.713	.807	.867	.944	.951	1.000
Sig. (1-tailed)	X1		.000	.000	.000	.000	.000	.000	.000
	X2	.000		.000	.000	.000	.000	.000	.000
	X3	.000	.000		.000	.000	.000	.000	.000
	X4	.000	.000	.000		.000	.000	.000	.000
	X5	.000	.000	.000	.000		.000	.000	.000
	X6	.000	.000	.000	.000	.000		.000	.000
	X7	.000	.000	.000	.000	.000	.000		.000
	X8	.000	.000	.000	.000	.000	.000	.000	

a. Determinant = 9.53E-007

KMO and bartlett test :

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.898
Bartlett's Test of Sphericity	Approx. Chi-Square	686.270
	df	28
	Sig.	.000

Communalities :

	Initial	Extraction
X1	1.000	.918
X2	1.000	.945
X3	1.000	.845
X4	1.000	.838
X5	1.000	.898
X6	1.000	.972
X7	1.000	.964
X8	1.000	.938

Extraction Method: Principal Component Analysis.

Total variance explained :

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	6.681	83.514	83.514	6.681	83.514	83.514	4.075	50.935	50.935
2	.637	7.968	91.483	.637	7.968	91.483	3.244	40.548	91.483
3	.232	2.896	94.379						
4	.213	2.662	97.041						
5	.096	1.195	98.235						
6	.081	1.015	99.251						
7	.048	.596	99.847						
8	.012	.153	100.000						

Extraction Method: Principal Component Analysis.

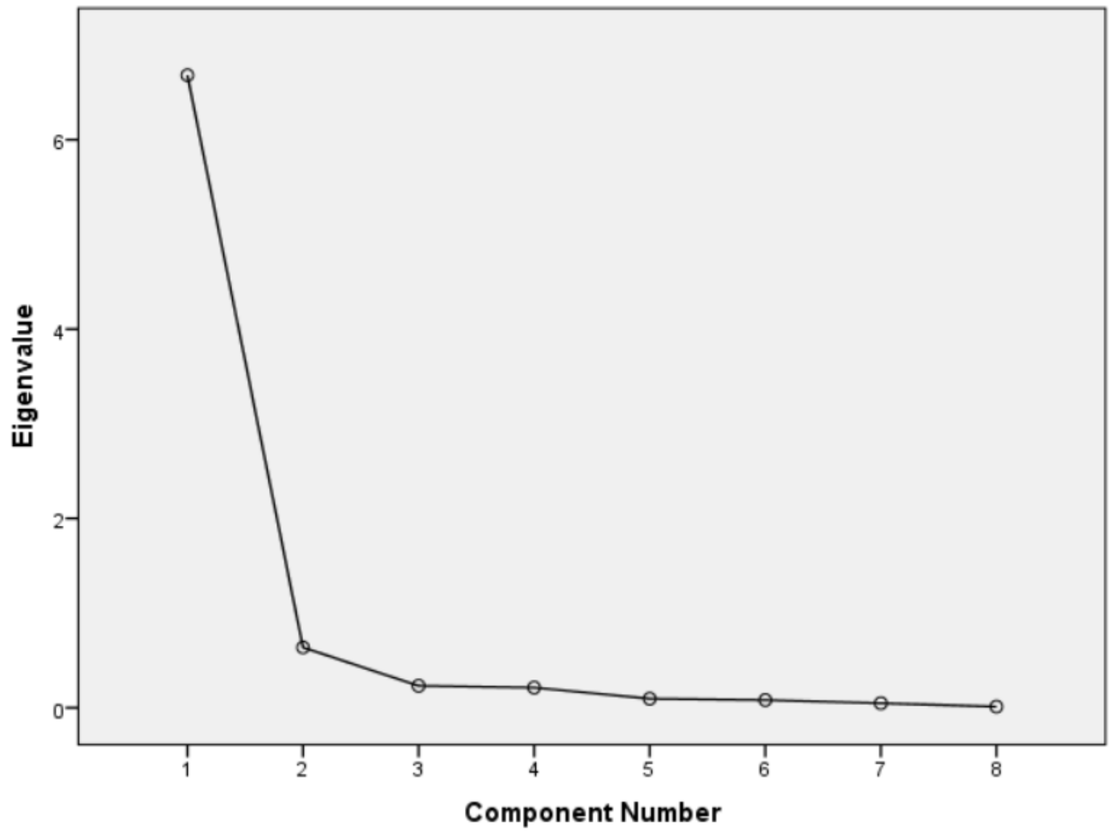
Rotated component matrix:

	Component	
	1	2
X8	.896	.367
X7	.892	.411
X6	.878	.449
X5	.781	.537
X4	.733	.548
X1	.376	.881
X2	.432	.871
X3	.479	.785

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.[a]

a. Rotation converged in 3 iterations.

Scree plot :



Part 6:

Conclusion

INFERENCE

Inferences

1 Segregating variables

On the basis of factor loadings of each factor, we see that X4 to X8 are the variables that define factor 1. Variables X1 to X3 define factor 2.

2 Domain knowledge

With the help of domain knowledge, short-distance running requires speed. Whereas, long-distance running requires endurance.

3 Relevant names

Since factor 1 represents long-distance races and factor 2 represents short-distance races, relevant names for the two factors can be Endurance and Speed.

Countries grouped on the basis of performance

On the basis of factor loadings of each country, two clusters can be formed. One with countries that perform well in long-distance races and another with better performance in short-distance races.

Long Distance Races	Short Distance Races
Argentina	Australia
Austria	Brazil
Belgium	Canada
Bermuda	Chile
China	Columbia
Cook Islands	Costa Rico
Czech Republic	Germany
Denmark	Great Britain
Dominician Republic	Greece
Finland	Guatemala
France	Hungary
Indonesia	India
Ireland	Israel
Kenya	Italy
Korea, South	Japan
Malaysia	Korea, North
Mauritius	Luxembourg
Mexico	Myanmar(Burma)
Netherlands	Philippines
New zealand	Poland
Norway	Portugal
Papua New Geinea	Russia
Romania	Taiwan
Samoa	Turkey
Singapore	U.S.A
Spain	
Sweden	
Switzerland	
Thailand	

Countries with similar performance

On the basis of difference in factor loadings, the top five countries having the least difference are found. These are the countries having similar performance in both kind of races.

France
China
India
Germany
Korea, South

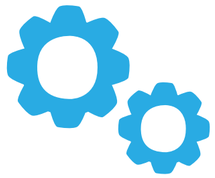
Part 7:

Limitations

DRAWBACKS

- Usually, it is recommended to keep the components with eigenvalues greater than 1. In this dataset, the factors that we considered had eigenvalues 6.68 and 0.64

- The dataset contained men track records of just 54 countries. This drawback has also been discussed as a disadvantage of MVA



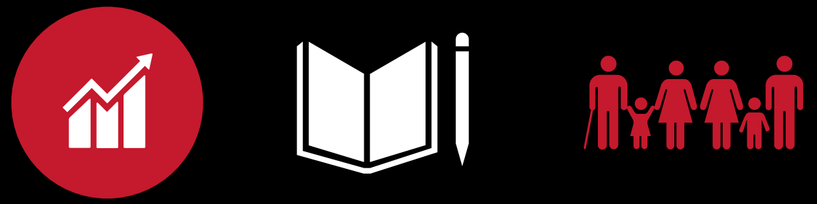
Part 8:

Future work

POSSIBLE STUDIES

- Similar data can be collected from countries all over the world to generalize results.

- By not limiting ourselves to races, data relating to different kinds of sport activities can be used to generate meaningful and effective results.



References

PCA using Python
- https://youtu.be/QdBy02ExhGI

PCA using R programming
- https://youtu.be/0Jp4gsfOLMs

PCA using SPSS
- https://www.youtube.com/watch?v=nkvdUUCudEw
- https://stats.idre.ucla.edu/spss/seminars/efa-spss/

Basics of PCA
- https://youtu.be/pmG4K79DUoI
- https://iq.opengenus.org/applications-of-pca/#:~:text=The%20principal%20application%20of%20PCA%20is%20dimension%20reduction.,many%20high%20dimensions%20is%20captured%20in%20fewer%20dimensions.

Data source
- IAAF/ATES Tracks and Field Statistics Handbook for the Heisinki 2005 Olympics. Courtesy of Ottavio Castellini.

Acknowledgement

The success and final outcome of this assignment required guidance from many people and we are extremely fortunate to have got this all along the completion of the project work.

We respect and thank Prof. Preeti Ravikiran for giving us the opportunity to learn and explore.

We are really grateful to Dr. Santosha C D for his efforts and guidance.

THANK YOU!

PREPARED BY -

APOORVA GUPTA

MO01

7629980943

NISTHA RULANIA

MO08

9929424763

SUHASI GOHIL

MO12

9831950055