FIFA World Cup group stage performance

A trophy and a football ball in a stadium

Description automatically generated

**S18825**

**Reboly Seevaratnam**

Introduction

Factor analysis is a statistical technique used to identify underlying factors or dimensions that explain the relationship between the observed variables. The main objective of Factor analysis is to reduce dimensionality by grouping variables that are highly correlated with each other into smaller number of factors. In this study we need to find the number of factors that are necessary to explain the data set. Once we find the number of factors, we need to perform hypothesis test to find out whether the selected factors are enough or not.

Originally, the data ‘FootBall\_data.csv ’ has been collected from 2018 FIFA World Cup Official App and was updated to “kaggle” website later. Data set describes about each and every aspect of the football matches took place in 2018 FIFA Group stages This data contains 14 variables, each containing 32 rows. I have selected only the numerical values of the given data set. Here I am going to perform the factor analysis techniques using principal component methods. As different values have different measurement units the data set was standardized. The description of the variables are given below

* "group"
* "team"
* "rank"
* "wins"
* "losses"
* "goals\_scored"
* "goal\_difference"
* "points"
* "matches\_played"
* "exp\_goal\_difference"
* "draws"
* "goals\_against"
* "expected\_goal\_scored"
* "exp\_goal\_difference\_per\_90"
* "points"
* "exp\_goal\_conceded"

Importance of Group Stage Performance

Determining Advancement: The primary objective of teams in the group stage is to secure qualification to the knockout stages of the tournament. Teams compete against each other within their respective groups, and the top teams from each group advance to the next round. Therefore, strong performances in the group stage are crucial for teams to progress further in the tournament.

Testing Tactics and Strategy: The group stage allows teams to test different tactics, formations, and strategies against a variety of opponents. Coaches use the group stage as an opportunity to assess the strengths and weaknesses of their team and make adjustments accordingly for the later stages of the tournament.

Methodology

Data Collection

The data ‘FootBall\_data.csv ’ has been collected from 2018 FIFA World Cup Official App and was updated to “kaggle” website later.

Data Preprocessing

Collected data was cleaned, standardized, and prepared for analysis, including handling missing values and outliers.

Factor Analysis

Identify underlying factors or latent variables that explain patterns of correlations among observed variables. Factor analysis is used to simplify and interpret complex data sets by reducing the number of variables to a smaller set of underlying factors.

Software and Tools

R is a widely used open-source programming language and environment for statistical computing and graphics. It offers a comprehensive suite of packages and functions for conducting factor analysis and other statistical analyses.

Analysis

Factor Extraction: Describe the process of extracting factors from the observed variables using factor analysis, including selecting the number of factors and interpreting factor loadings.

Factor Rotation: Discuss any rotation techniques used to enhance the interpretability of the factors and simplify the factor structure.

Interpretation of Results

**Eigen values**

**A close-up of a number

Description automatically generated**

When we consider the above Eigen Values, there are six variables with eigen value greater than one. Therefore, by using the Eigen values we can state that six factors are sufficient to explain the data set.

**PROPORTION OF VARIANCE EXPLAINED**

A close up of a number

Description automatically generated

In Here we Add about Proportion of each variable variance explained separately.

**Eigen Vector**

A screenshot of a computer code

Description automatically generated

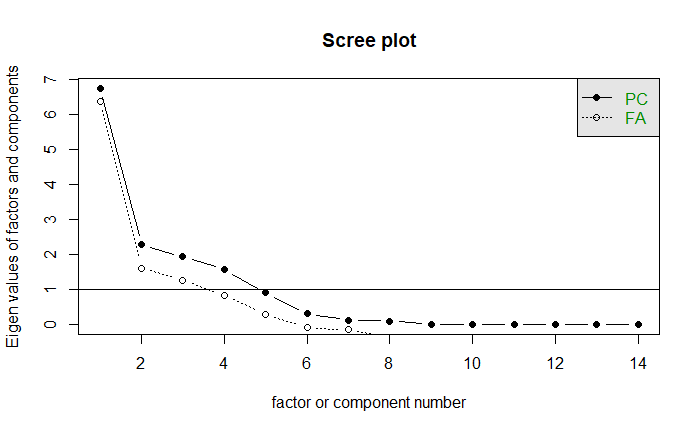
This is an eigen Vector that we got using this analysis.

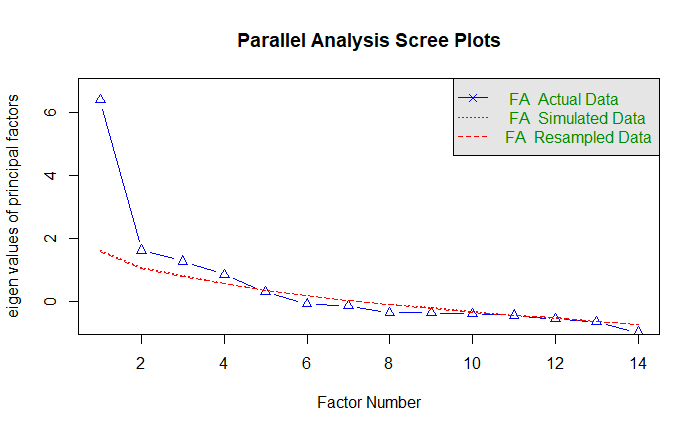
**Scree plots**A graph with numbers and lines

Description automatically generated

Using above scree plot we can see that the Elbow value of this graph is 4.

So we Take number of factor is 4.



**Factor Diagrams**A diagram of a factor analysis

Description automatically generated

**Factor Loadings**

A screenshot of a computer

Description automatically generated

**Communalities**

|  |
| --- |
|  |
|  |  |
| A screenshot of a computer  Description automatically generated |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

Conclusion

\* Variables such as "group", "wins", "losses", "goal\_difference", and "points" have high communalities, indicating that a large proportion of the variance in these variables is accounted for by the underlying factors extracted in the factor analysis.

\* Variables such as "rank", "goals\_against", and "goal\_difference" have moderate communalities, indicating that a substantial but not complete proportion of the variance in these variables is explained by the factors.

\* Variables such as "draws", "goals\_scored", and "goals\_against" have relatively low communalities, suggesting that a smaller proportion of the variance in these variables is explained by the factors.

\* High communalities indicate strong relationships between variables and underlying factors, supporting the interpretation of factor loadings and the overall factor structure.

\* Low communalities may warrant further investigation to determine the reasons for their poor performance and whether adjustments to the factor analysis model or data preprocessing are necessary.

Dataset

https//www.kaggle.com/datasets/mathan/fifa-2018-match-statistics

**CODES**

library(ggplot2)

library(data.table)

library(tidyverse)

library(janitor)

library(factoextra)

library(psych)

library(corrplot)

Football\_Data=read\_csv("../Football\_Analysis/FootBall\_data.csv")

head(Football\_Data)

Football\_Data[is.na(data)] = 0

describe(Football\_Data)

colnames(Football\_Data)

nrow(Football\_Data)

ncol(Football\_Data)

numeric\_data=Football\_Data[,sapply(Football\_Data,is.numeric)]

head(numeric\_data,5)

numeric\_data\_subset=numeric\_data[,!(names(numeric\_data)=="matches\_played")]

head(numeric\_data\_subset)

cor(numeric\_data\_subset,method = "spearman")

apply(numeric\_data\_subset,2,mean)

#STANDARDIZE DATASET

data\_std <- apply(numeric\_data\_subset,2,scale)

#COORELATION MATRIX

cor\_matrix = cor(data\_std)

#COVARIANCE MATRIX

cov\_matrix = cov(data\_std)

#KMO VALUE

KMO(data\_std)

apply(numeric\_data\_subset,2,mean)

data.cov = cov(data\_std)

data.eigen = eigen(data.cov)

data.eigen$values

data.eigen$vectors

#PROPORTION OF VARIANCE EXPLAINED

PVE = data.eigen$values/sum(data.eigen$values)

PVE

#SCREE PLOT

scree(data\_std)

plot(data.eigen$values,main = "Scree Plot",xlab = "Factors",ylab = "Eigen

Values",pch=20,col = "blue")+lines(data.eigen$values,col ="red")+abline(h=1,col

="green",lty = 2)

#PARALLEL ANALYSIS

fa.parallel(data\_std,fm = "pa",fa = "fa")

fourfactors = fa(r=data\_std,nfactors = 4,rotate = "none",fm="pa")

print(fourfactors)

#FACTOR ANALYSIS USING PRINCIPAL COMPONENTS METHOD WITHOUT ROTATION

data\_std\_fa\_pc <- fa(data.cov,nfactors = 4,rotate = "none",n.obs = 128,covar =

TRUE,fm = "pa")

data\_std\_fa\_pc

fa.diagram(data\_std\_fa\_pc)

#FACTOR LOADINGS

data\_std\_fa\_pc$loadings

unroated\_pc\_loadings <- as.data.frame(unclass(data\_std\_fa\_pc$loadings))

unroated\_pc\_loadings

unroated\_pc\_com <- as.data.frame(unclass(data\_std\_fa\_pc$communality))

unroated\_pc\_com

#FACTOR ANALYSIS USING MAXIMUM LIKELIHOOD METHOD WITH ROTATION

data\_std\_fa\_ml <- fa(data.cov,nfactors = 4,rotate = "none",n.obs = 128,covar =

TRUE,fm = "pa")

data\_std\_fa\_ml

unroated\_ml\_loadings <- as.data.frame(unclass(data\_std\_fa\_pc$loadings))

unroated\_ml\_loadings

unroated\_ml\_com <- as.data.frame(unclass(data\_std\_fa\_pc$communality))

unroated\_ml\_com

#FACTOR ANALYSIS USING PRINCIPAL COMPONENTS METHOD WITH ROTATION

data\_std\_pc\_rotate <- fa(data.cov,nfactors = 4,rotate = "varimax",n.obs =

128,covar = TRUE,fm = "pa")

data\_std\_pc\_rotate

data\_std\_pc\_rotate$loadings

rotated\_pc\_loadings <- as.data.frame(unclass(data\_std\_pc\_rotate$loadings))

rotated\_pc\_loadings

rotated\_pc\_com<- as.data.frame(unclass(data\_std\_pc\_rotate$communality))

rotated\_pc\_com

DATASET

A screenshot of a calendar

Description automatically generatedA screenshot of a computer

Description automatically generated