Contents

Abstract

For many years now, countries' relative development has been focused on individual variables such as Gross Domestic Product (GDP). The widely referenced Human Development Index (HDI), introduced in 1990 in an attempt to broaden the focus from GDP alone, is only composed of three "core" variables. The main purpose of this report is to identify and expand on these variables using Principal Component Analysis (PCA).Furthermore , PCA conducted on the data ,show the results are mostly in agreement with the definition of HDI. The report also highlights , the merits of PCA to reveal clusters formed between countries and using it as a precessor to conducting cluster analysis.

Introduction

The data for the project comes from the World Development Indicators, which is a database compiled by the World Bank from various international sources containing accurate and up-to-date time series concerning topics such as education, health, and economics. The database is updated quarterly and covers more than two hundred economies. From the original database, this project will be based on a selection of 19 variables regarding 196 countries from an initial 22 variables .This is due to some variables having a large number of observations as missing as well as omission of these variables provides better interpretation of the results from PCA . The variables selected can be grouped into topics, in this case: Agriculture and Food (5), Demographics (3), Technology (3), Health (2), Economy (2), Education (1), Transport (1), Tourism (1) and Pollution (1)1. As this data takes time to collect, verify and publish, the chosen reference year is 2013 since this is the most recent year with most of the data available.

As mentioned above ,despite the data from 2013 being the most complete, it is still missing some data for certain countries and variables. An example of this is the variable Tax Revenue where around 40% of the data is not available. However, this does not appear to be a big issue with the data set as there are only a few variables with this volume of missing values. To overcome this , the Random Forest package is used to to impute these values and further information regarding these variables are obtained using information from previous years.

Using principal Component Analysis, the loadings are interpreted and possible trends explained for the high loading variables. In addition, due to large differences in variable units, the analysis is carried after scaling the data. As a final extension , the report looks into identifying if the natural patterns in the features of the data can be explained by the addition of external variables.

Data Preperation

Preparing the data

The original data set contained 196 observations with 22 variables .The variables were renamed for ease of use , for example : “Internet users (per 100 people)” renamed as “InternetUsersPer100”.The full list of the original variables ,abbreviations and the description is provided in the Appendix. Furthermore , several countries that had been aggregated has also been excluded in the final data set. The countries in particular no order are :Caribbean Small states ,Fragile and conflict affected situations,North America,Other small states,Pacific island small states and Small states

Missing Data and Imputation Methods :

The “Amelia Package” in R can be used to show the severity of “missingness” between the variables in the data

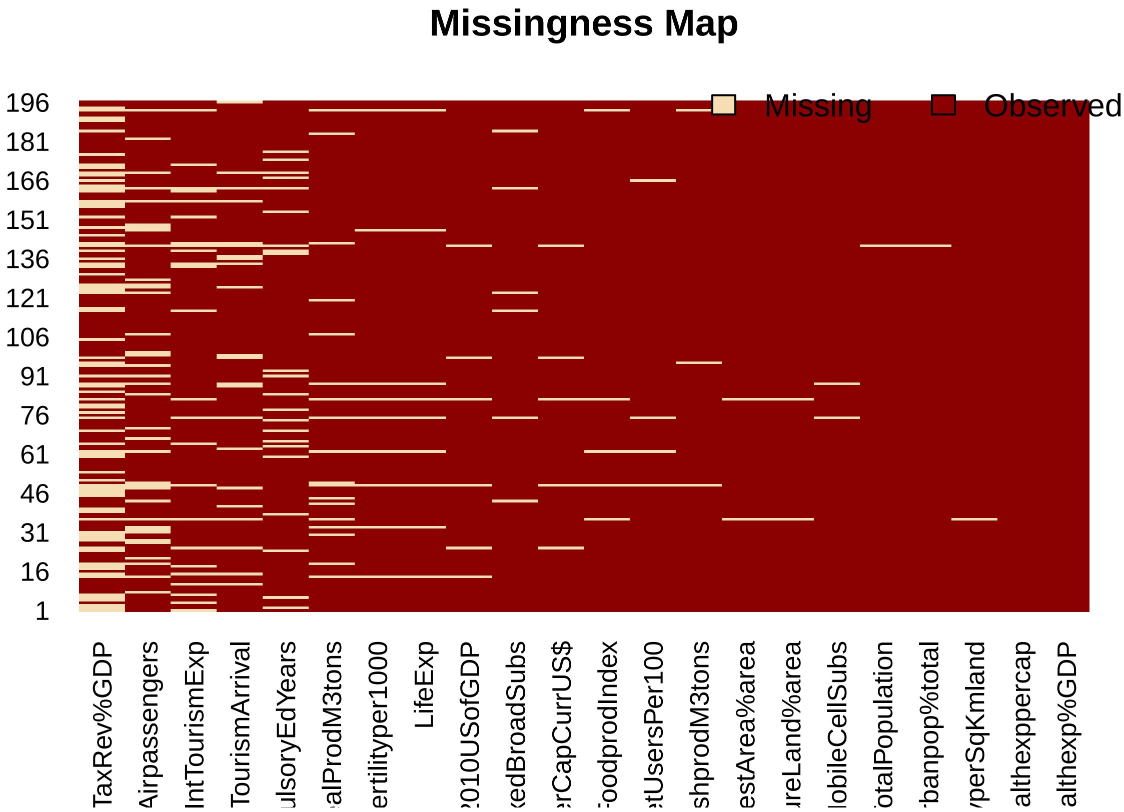

Figure 1: Missingness map of the orginal datset

The variable Tax Rev%GDP , has approximately 40% of the rows missing , and is replaced by the variable “Unemployment rate” . The final data set has therefore ,190 observations and the observation with the most missing values has 29%.

The “Random Forest” package in R is used for imputing the missing values . Records with missing values in the variable are imputed by random draws from independent normal distributions centered on conditional means predicted using random forest. Random forest fits each tree to a different bootstrap sample of the data and aggregates the results (Shah et al. 2014) . And hence the imputation method is carried out post scaling/transformations of the data(satisfying the normality criteria) .

Transformation of Variables:

An initial look at the histograms of the different variables , identifies large skewness in many variables thereby violating the normality condition for the PCA . Plotting the histograms for those variables :

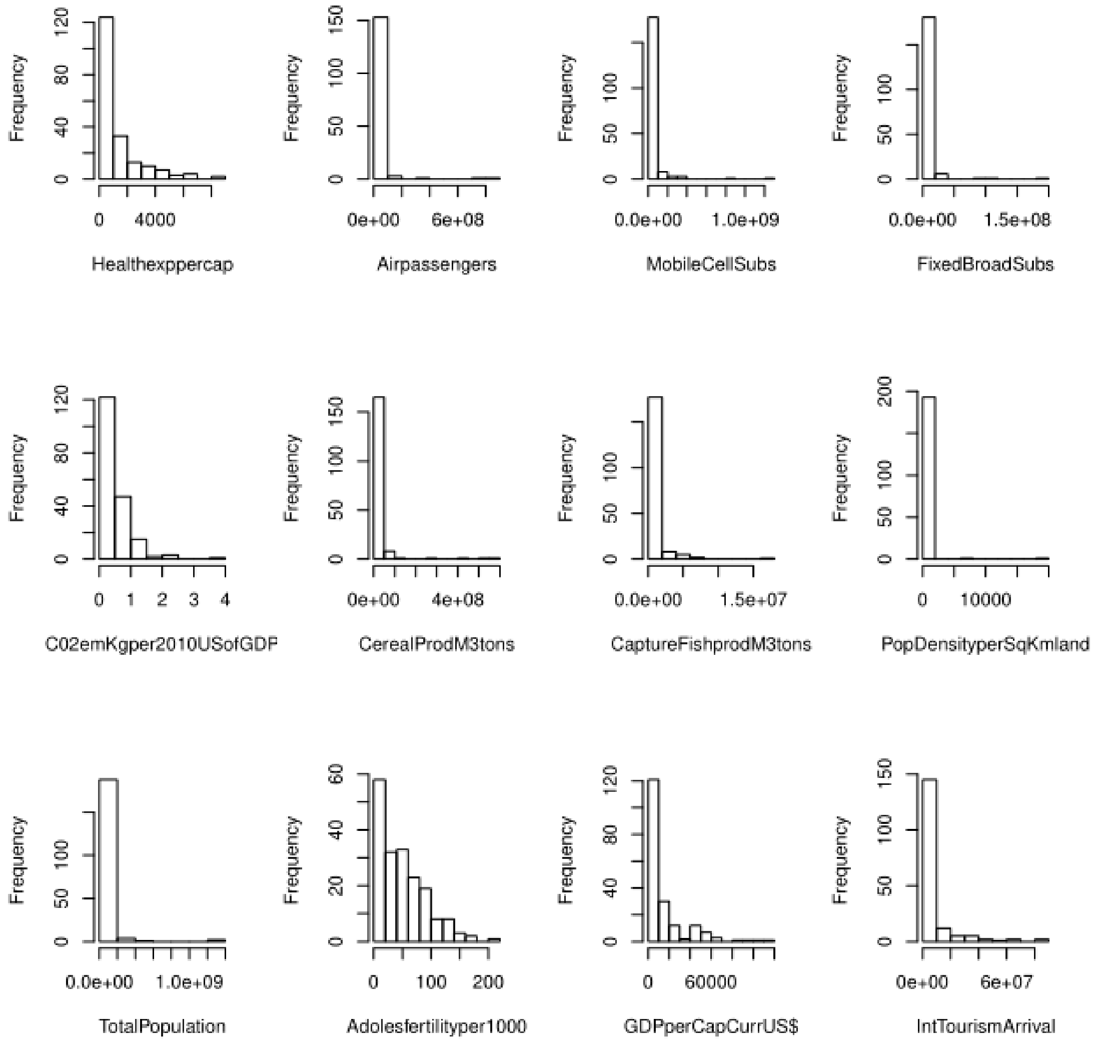

Figure 2: Histograms of variables violating normality assumption

As seen in Fig 2, A large no of a variables violate the normality condition and therefore require to be transformed to produce better results : A log base e transformation is applied to the following Healthexppercapita, CO2emKgper2010USofGDP, PopDensityperSqKmland, Adolesfertilityper1000,GDPperCapCurrUS., Unemployment.rate, FixedBroadSubspertotpopulation, CaptureFishprodM3tonspertotpopulation, IntTourismArrivalpertotpopulation were all log transformed

As a side note,the variables CerealProdM3tonspertotpopulation and Airpassengers population have large count's of zeros , and transforming by log(1+x) did not produce any better results in terms of loadings and hence left untransformed .

Scaling of Variables

From Fig 2, we see the "TotalPopulationvariable" has large variation in data with the Minimum value at $9.876*10^3$,the Maximum value at $1.357*10^9$, and the median at $8.060*10^6$. Furthermore looking at our initial data set , we see a number of variables that are dependent on the TotalPopulation. And hence the variables: Airpassengers,MobileCellSubs,FixedBroadSubs,CerealProdM3tons, CaptureFishprodM3tons,IntTourismArrival are all divided by the TotalPopulaton and re-scaled .

Furthermore the variable ,"IntTourismExp" is excluded in the final analysis due to ambiguous nature of the variable definition .

Defining External Variables

Addition of external variables to the original data set allows to color-code samples by external variables to see if samples sort according to those variables . This is particularly useful in PCA , where natural patterns or formations can be observed and the validity of the results can be assessed .

With this in mind ,the external variables chosen are in categorical nature with some discretised :

Table 1: Definition of External Variables added

Categorical Variable Alias	Definition	Discretised
Continent	Name of the Continent to which the country belongs	No
Olympics	Has the country ever hosted the Olympics	Yes
Nuclear	Does the country have nuclear weapons	Yes
Coastline	Is the country surrounded by water / has a coast	Yes

A detailed table of all the variables ,definitions and corresponding abbreviations are provided in the appendix .

Method

Explanatory Data Analysis

Principal Component Analysis as a procedure converts a set of possibly correlated variables into a set of values of linearly uncorrelated variables , i.e reduces the dimension of the original data set. An initial step is therefore , to look at the correlation of the original data set before and after scaling .

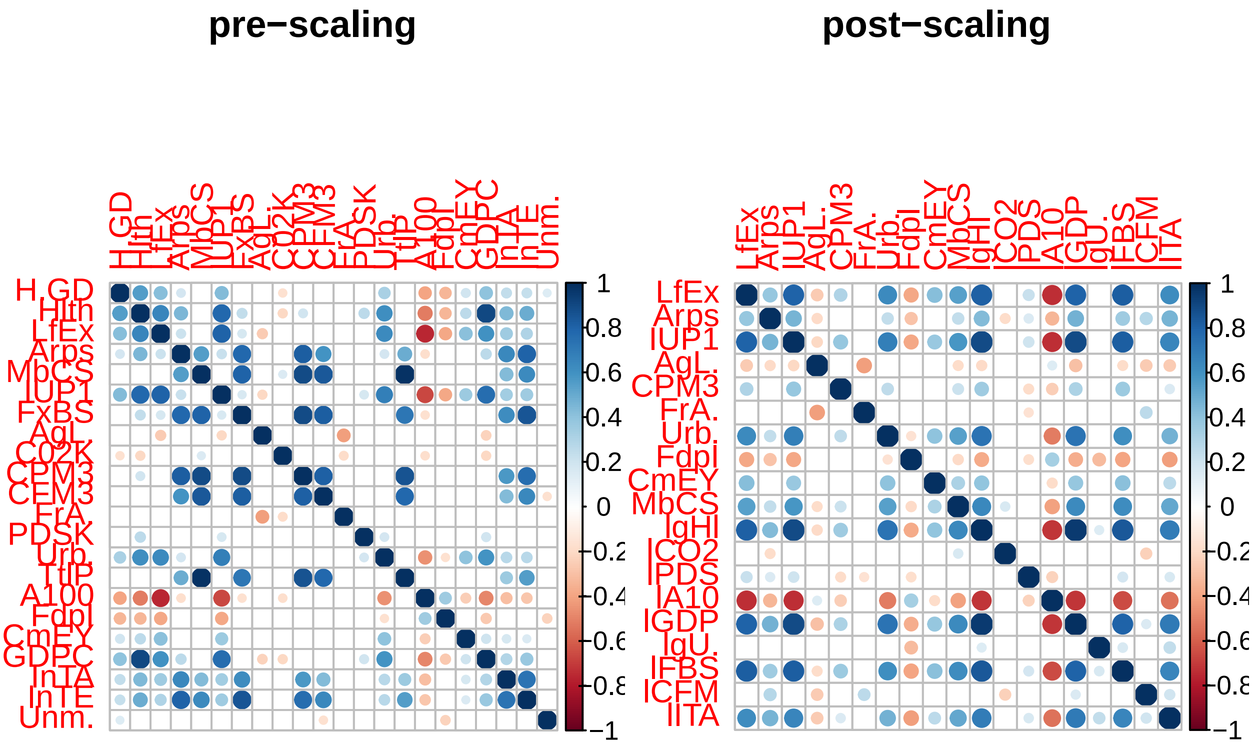

Figure 3: Correlation plots pre and post transformating-scaling of the variables

In Fig 3, Insignificant correlations between variables are left blank .Post-transformations and scaling , the correlations appear stronger.As shown later , the 1st Principal component can be formed from the "lA10" strongly correlated with variables such as "LifeExp","IUP1","Urb","MBCS" etc.This relationship becomes less clear in the correlation matrix pre-scaling , and results loadings difficult to interpret.

Choice of correlation matrix using the princomp function in R

In PCA, the choice of correlation and co variance matrices give different results for the loadings and scores . With the data set in mind , a natural route is to use the correlation matrix due to the difference in standard deviations even after scaling. Running the PCA with and without a correlation matrix , provides further confirmation.The resulting "Scree and Bi- plots" are given in figures 4,5.

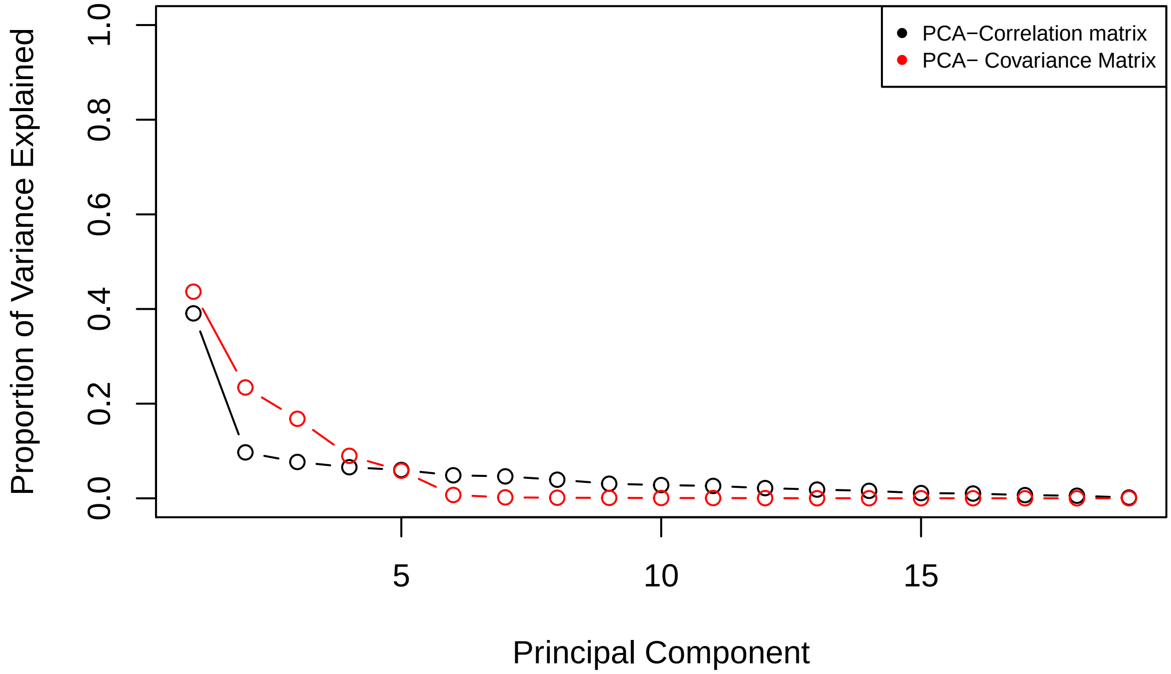

Figure 4: Scree plots of with and without Correlation Matrix

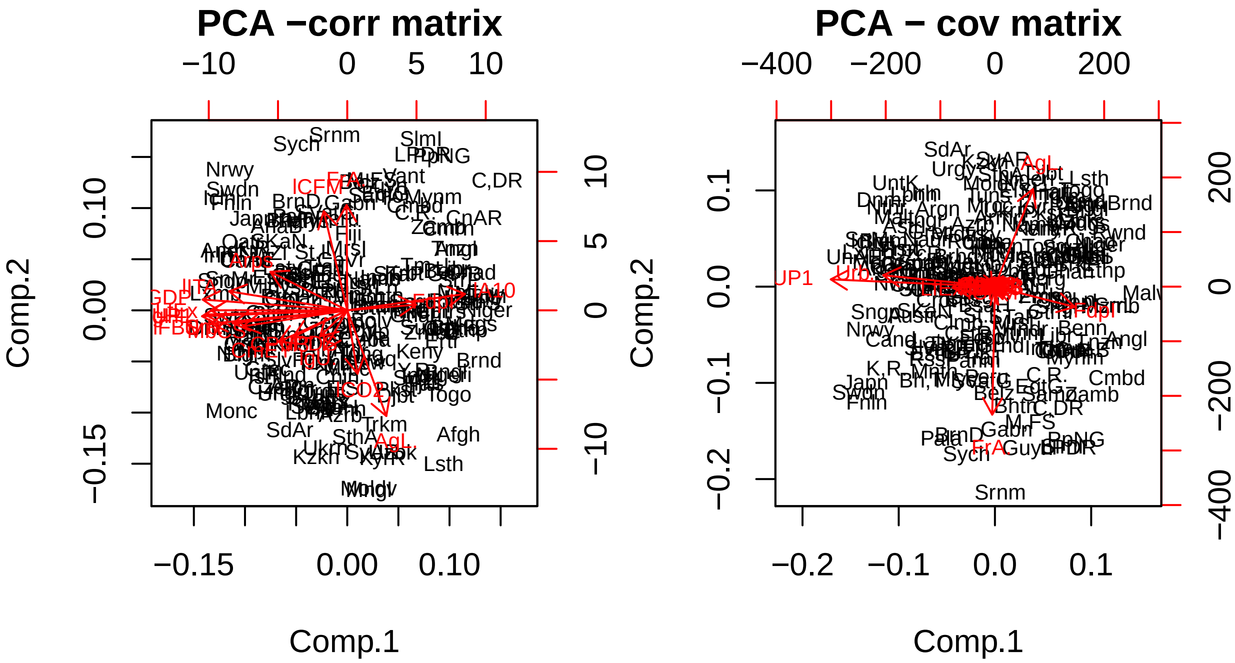

Figure 5: Bi-plots of with a Correlation and Covariance Matrix

The scree plots suggest , the correlation matrix to be the better choice as fewer components explain the variance in data. However, the bi-plots reveal the variables dominating the PC1 and PC2 components , are dominated by those variables (IUP1,Urb,Agl,FrA) which have naturally high units and large variances , thereby a correlation matrix is used to standardize the data .

PCA on scaled data

Selecting the Principal Components:

As an initial step , the most important PC components is selected using Kaiser's Criterion (Gentleman et al. 2008) ,defined as retaining the components that has eigenvalues greater than the mean eigenvalue . However, the 6th component is also included in the final loadings , due to the findings in the Results section.

The loadings obtained are then interpreted with the addition of external variables. Loading's provide which variables contribute to the 5 principal components . "Large"" Positive Loading's , indicate a variable and principal component to be strongly positively correlated , and a negative one the reverse. Furthermore, Mardia's condition is used to differentiate between "large" and "small" loadings (Gentleman et al. 2008) .

Results

Interpretation of Loadings :

PC1 Loadings:

As in Table 1.1, 9 variables are deemed important to the 1st principal component . The 1st component explains approximately 39.04% of the variance in data and separates countries with High Health Expenditure/GDP per Capita / Internetusers.. etc and low adolescent fertility rate and similarly low Health Exp/GDP per capita… and high adolescent fertility rate. The above also ties in well with the 2 of the 3 indicators used by the UNDP in developing the "Human Development Index". In particular GDP per Capita , LifeExp have high loadings in the first component , although the report finds HealthexpperCapita to be a more useful variable than LifeExp.

PC2 Loadings :

The second component is largely concerned with "Land and food" of the countries, in combination with PC1 , it explains 48.77% of the variance in the data. On the 2nd principal component , countries with large agriculture land area have lower number of fishes captured , indicating these countries dependency on farming or viceversa.

PC3 & PC4 Loadings :

The 3rd and 4th components is majorly concerned with the " population density " and “”Unemployment rate" of the countries

PC5 Loadings:

On the 5th component , a contrast between "Cereal Productions" and "CO2emission" is identified , as previous the component, will separate between countries for high Cereal Production and C02 emissions with lower Cereal and C02 emissions.

PC6 Loadings:

The 6 components chosen , explains approximately 74% of the variance in the data. The 3rd variable "Education" part of the "Human Development Index" only appears on the 6th component .A possible explanation ,is the data collected for "Compulsory" only looks at the duration for the "compulsory" years spend in education , while "Education Index" takes into account how long a child spends in education as well as the number of years a child can be expected to spend in a given level of education .Furthermore , " CompulsoryEdYears" ranks 5th in "missingness" of data with 11.22% of the observations missing and hence imputation method needs to be revisited.

Interpretation of Scores using the Bi-Plot

The interpretation of the loadings can now be confirmed by taking a closer look at the resulting bi-plot :

As suggested before , we see a separation on the PC1 axis ,between developing countries (with low GDP , low Life Exp , low internet users.. and high adolescent Fertility rates) on the right and developed countries on the left with (High GDP , High Life Exp ,High internet users.. and low adolescent fertility rates) .

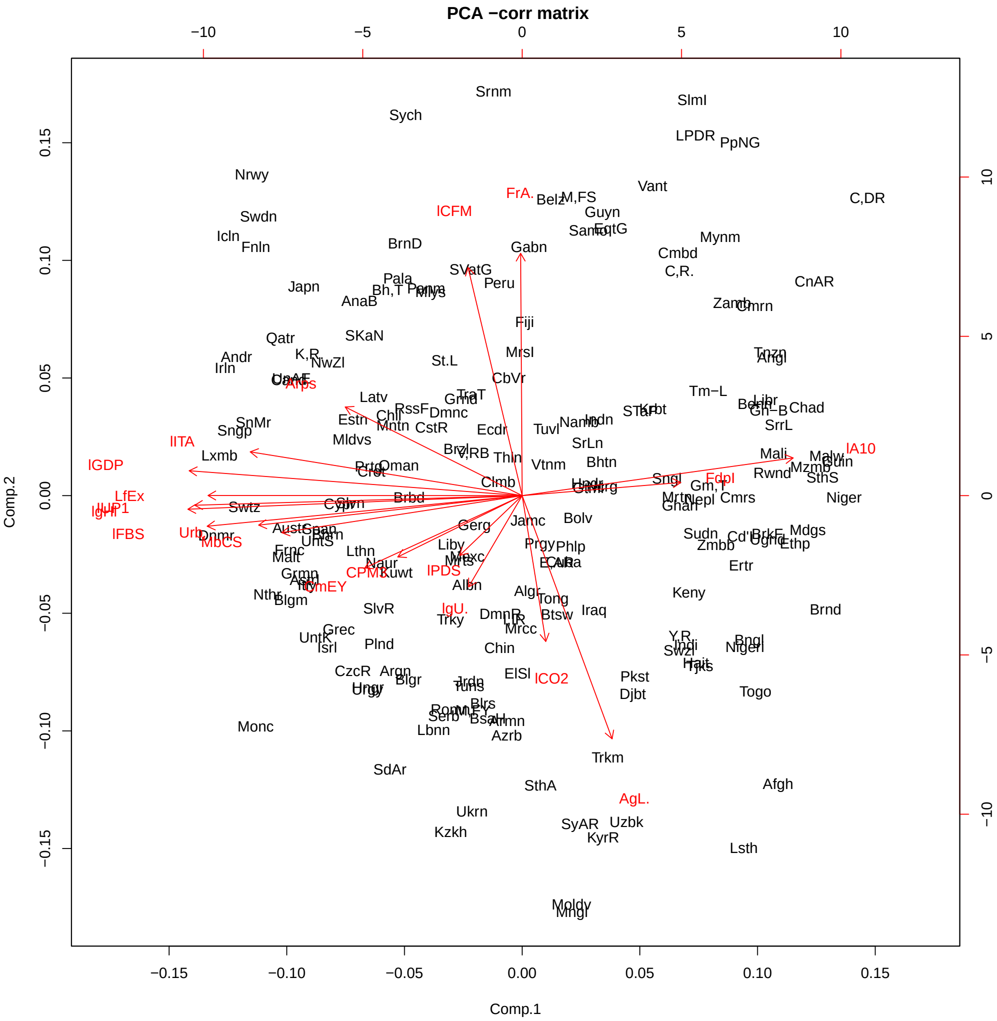

Figure 6: Biplot on the correlation matrix

And similarly we see a separation on the PC2 axis ,with countries placed at the end of vector “ICFM”i.e Capture Fisheries such as “St. Lucia St. Vincent and the Grenadines”“-”Svat" and “Fiji” being islands highly dependent on Fish and less on Agricultural Land.

Addition of External Variables

External variables labelled as “Continents”"

To confirm the findings in the previous , a plot can be obtained for the scores in the first two principal components with values for “GDPperCap” and “InternetUSersper100” coded using a color scale and labels added for the continents to which the countries belong to .

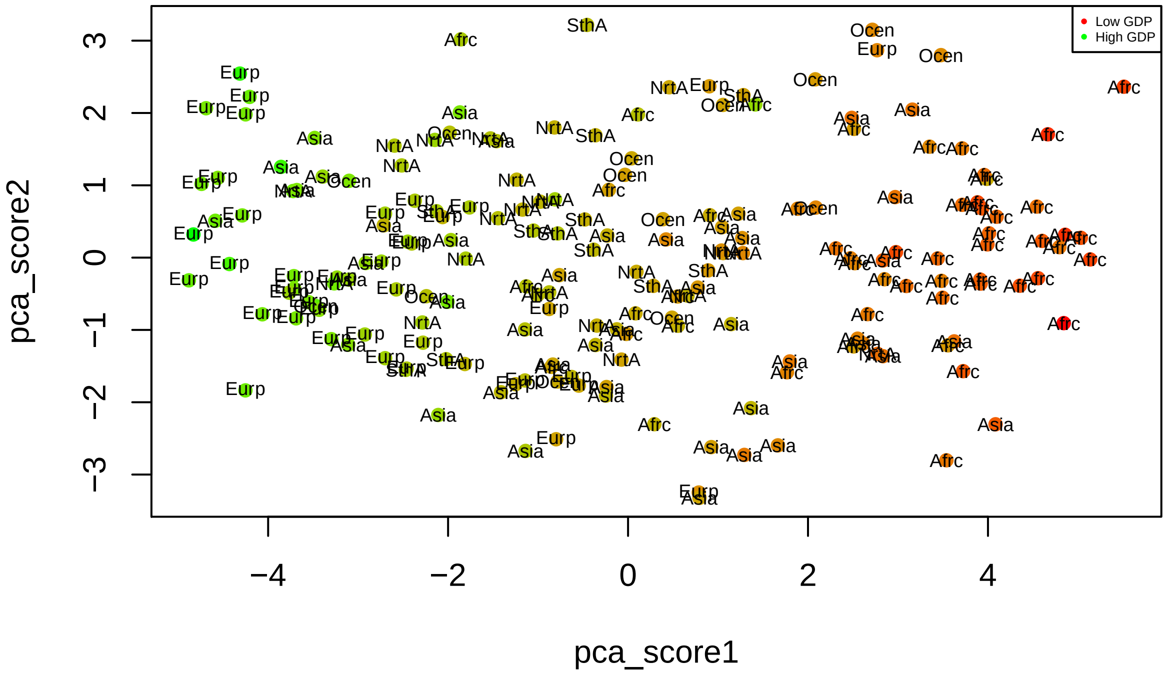

Figure 7: Plot of Scores with external variables added for continents and GDP on a color scale

Fig 7, countries in the African continent have a lower GDP compared to the other end of the spectrum ,with European Continents on the left (high) end of the spectrum.The PC1 axis is able to separate between developing and developed countries

External Variables Labelled as “”coastline "

The plot in Fig8, confirms how PC2 separates countries with high positive scores due to “CaptureFisheries” are the countries with a coastline and vice versa

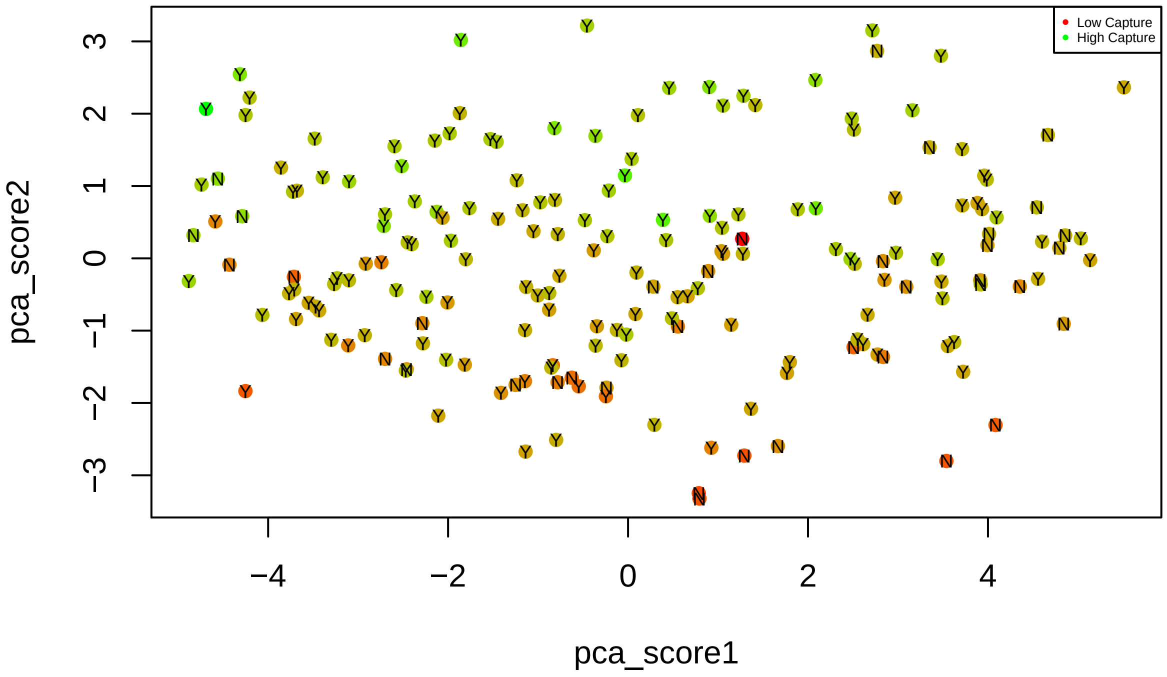

Figure 8: Plot of Scores with external variables added for coastline and variable ’CaptureFishproduction’on a color scale

Conclusion & FurtherWork

The results , show Principal Component Analysis , is able to provide a distinction between countries based on the different variables .Furthermore reducing , the dimensionality of the original 19 variables to 6 main components, PCA correctly identifies countries based on their inherent characteristics. For example ,in Fig 6 ,the top left hand side of the plot, the PC1 and PC2 components reveals clusters for "Norway","Sweden","Finland" and "Iceland" i.e Nordic Countries .And hence , a natural step to proceed is to run a k-means and hierarchical cluster analysis on the PCA scores to identify these clusters. The report also finds HDI developed by the UNDP are based on variables identified by the PCA as the most important ,except for "Education".To improve upon this , any further work should explore the variables chosen more thoroughly, such that a scoring for each of the countries can be developed based on the weights of the important variables (weighted PCA) (Lai 2003) .

Appendix

Table 1.1:Loadings Based on Mardia's Criterion

- **PC1 Loadings on MArdia**:

Table 2: loadings of the first 6 components on Mardia's criterion

	Mardiaspc1laodings
LifeExp	-0.328
InternetUsersPer100	-0.3418
Urbanpop.total	-0.2751
MobileCellSubspertotpopulation	-0.251
logHealthexppercapita	-0.349
logAdolesfertilityper1000	0.283
logGDPperCapCurrUS.	-0.3477
logFixedBroadSubspertotpopulation	-0.3289
logIntTourismArrivalpertotpopulation	-0.2839

- **PC2 Loadings on MArdia**:

	Mardiaspc2laodings
AgricultureLand.area	-0.5099
ForestArea.area	0.508
logCaptureFishprodM3tonspertotpopulation	0.4783

- **PC3 Loadings on MArdia**:

	Mardiapc3loadings
logPopDensityperSqKmland	0.5644

- **PC4 Loadings on MArdia**:

	Mardiapc4loadings
logUnemployment.rate	0.765

- **PC5 Loadings on MArdia**:

	Mardiapc5loadings
CerealProdM3tonspertotpopulation	-0.4489
logCO2emKgper2010USofGDP	0.6164

- **PC6 Loadings on MArdia**:

	Mardiapc6loadings
CompulsoryEdYears	-0.754

	Mardiapc6loadings

Table 1.2 : Variable Abbreviations

	abbreviatedvariable
LifeExp	LfEx
Airpassengerspertotpopulation	Arps
InternetUsersPer100	IUP1
AgricultureLand.area	AgL.
CerealProdM3tonspertotpopulation	CPM3
ForestArea.area	FrA.
Urbanpop.total	Urb.
FoodprodIndex	FdpI
CompulsoryEdYears	CmEY
MobileCellSubspertotpopulation	MbCS
logHealthexppercapita	lgHl
logCO2emKgper2010USofGDP	lCO2
logPopDensityperSqKmland	lPDS
logAdolesfertilityper1000	lA10
logGDPperCapCurrUS.	lGDP
logUnemployment.rate	lgU.
logFixedBroadSubspertotpopulation	lFBS
logCaptureFishprodM3tonspertotpopulation	lCFM
logIntTourismArrivalpertotpopulation	lITA
Olympics	Olym
Nuclear	Nclr
Coastline	Cstl
Continent	Cntn

1.3: R Code

```r
require(Amelia)
worldbank<- read.csv("Worldbank (1).csv")


#Rename variables which are very long
#names(worldbank)
names(worldbank)[names(worldbank) == 'Health.expenditure.
                 .public....of.GDP...SH.XPD.PUBL.ZS.'] <- 'Healthexp%GDP'
names(worldbank)[names(worldbank) == 'Health.expenditure.per.capita..PPP..constant.2011.inter
                 national.....SH.XPD.PCAP.PP.KD.'] <- 'Healthexppercap'
names(worldbank)[names(worldbank) == 'Life.expectancy.at
                 .birth..total..years...SP.DYN.LE00.IN.'] <- 'LifeExp'
names(worldbank)[names(worldbank) == 'Air.transport..
                 passengers.carried..IS.AIR.PSGR.'] <- 'Airpassengers'
names(worldbank)[names(worldbank) == 'Mobile.cellular
                 .subscriptions..IT.CEL.SETS.'] <- 'MobileCellSubs'
names(worldbank)[names(worldbank) == 'Internet.u
                 sers..per.100.people...IT.NET.USER.P2.'] <- 'InternetUsersPer100'
names(worldbank)[names(worldbank) == 'Fixed.broadba
                 nd.subscriptions..IT.NET.BBND.'] <- 'FixedBroadSubs'
names(worldbank)[names(worldbank) == 'Agricultural.
                 land....of.land.area...AG.LND.AGRI.ZS.'] <- 'AgricultureLand%area'
names(worldbank)[names(worldbank) == 'CO2.emissio
                 ns..kg.per.2010.US..of.GDP...EN.ATM.CO2E.KD.GD.'] <- 'CO2emKgper2010USofGDP'
names(worldbank)[names(worldbank) == 'Cereal.prod
                 uction..metric.tons...AG.PRD.CREL.MT.'] <- 'CerealProdM3tons'
names(worldbank)[names(worldbank) == 'Capture.fisheries
                 .production..metric.tons...ER.FSH.CAPT.MT.'] <- 'CaptureFishprodM3tons'
names(worldbank)[names(worldbank) == 'Forest.area...
                 .of.land.area...AG.LND.FRST.ZS.'] <- 'ForestArea%area'
names(worldbank)[names(worldbank) == 'Population.dens
                 ity..people.per.sq..km.of.land.area...EN.POP.DNST.'] <- 'PopDensityperSqKmland'
names(worldbank)[names(worldbank) == 'Urban.populati
                 on....of.total...SP.URB.TOTL.IN.ZS.'] <- 'Urbanpop%total'
names(worldbank)[names(worldbank) == 'Population..tota
                 l..SP.POP.TOTL.'] <- 'TotalPopulation'
names(worldbank)[names(worldbank) == 'Adolescent.fertility.rate..births.per.1.000.women.age
                 s.15.19...SP.ADO.TFRT.'] <- 'Adolesfertilityper1000'
names(worldbank)[names(worldbank) == 'Food.productio
                 n.index..2004.2006...100...AG.PRD.FOOD.XD.'] <- 'FoodprodIndex'
names(worldbank)[names(worldbank) == 'Duration.of.comp
                 ulsory.education..years...SE.COM.DURS.'] <- 'CompulsoryEdYears'
names(worldbank)[names(worldbank) == 'Tax.revenue....
                 of.GDP...GC.TAX.TOTL.GD.ZS.'] <- 'TaxRev%GDP'
names(worldbank)[names(worldbank) == 'GDP.per.capita.
                 .current.US....NY.GDP.PCAP.CD.'] <- 'GDPperCapCurrUS$'
names(worldbank)[names(worldbank) == 'International.
                 tourism..number.of.arrivals..ST.INT.ARVL.'] <- 'IntTourismArrival'
names(worldbank)[names(worldbank) == 'International.tourism..expenditures.for.travel.items
                 ..current.US....ST.INT.TVLX.CD.'] <- 'IntTourismExp'
worldbank<- worldbank[,-27]
```

```
worldbank<-worldbank[,5:length(worldbank)]


missmap(worldbank)
#Check current state of missing values
#Check current state of missing values
invisible(sapply(worldbank, function(x) sum(is.na(x))/nrow(worldbank)*100))

##plot the histograms
par(mfrow=c(2, 4))

worldbank$`TaxRev%GDP`<-NULL
worldbank$`Healthexp%GDP`<-NULL
worldbank$FoodprodIndex<-NULL
worldbank$`Urbanpop%total`<-NULL
worldbank$`AgricultureLand%area`<-NULL
worldbank$CompulsoryEdYears<-NULL
worldbank$`ForestArea%area`<-NULL
worldbank$InternetUsersPer100<-NULL
worldbank$LifeExp<-NULL

for (i in 1:length(worldbank)){
  hist(worldbank[,i],xlab=paste( names(worldbank)[i]),main=NULL)}






par(mfrow=c(1,2))
unscaleddata<- read.csv("C:/Users/robin/Dropbox/Applied Multivariate Analysis/project/dataset.csv")


unscaleddata$Continent<-as.character(unscaleddata$Continent)

## Change name of the Australian COntinent to Oceania :
index<-which(unscaleddata$Continent=="Australia")
unscaleddata[index,length(unscaleddata)]<-"Oceania"


unscaleddata$Rule.Type<-NULL
unscaleddata$Flag.Colours<-NULL
unscaleddata$Flag.Colours.Raw<-NULL
unscaleddata$Official.Languages<-NULL
unscaleddata$Official.Languages.Raw<-NULL

## Assign the countries to the rownames
## assign the row names

unscaleddata_nocat<-unscaleddata[,3:24]

## Look at correlation plots instead
library(corrplot)
corr_data<-unscaleddata_nocat
```

```
colnames(corr_data)<-abbreviate(colnames(corr_data))
#corrplot(cor(corr_data), method="ellipse")


cor.mtest <- function(mat, conf.level = 0.95){
  mat <- as.matrix(mat)
  n <- ncol(mat)
  p.mat <- lowCI.mat <- uppCI.mat <- matrix(NA, n, n)
  diag(p.mat) <- 0
  diag(lowCI.mat) <- diag(uppCI.mat) <- 1
  for(i in 1:(n-1)){
    for(j in (i+1):n){
      tmp <- cor.test(mat[,i], mat[,j], conf.level = conf.level)
      p.mat[i,j] <- p.mat[j,i] <- tmp$p.value
      lowCI.mat[i,j] <- lowCI.mat[j,i] <- tmp$conf.int[1]
      uppCI.mat[i,j] <- uppCI.mat[j,i] <- tmp$conf.int[2]
    }
  }
  return(list(p.mat, lowCI.mat, uppCI.mat))
}

res1 <- cor.mtest(corr_data,0.95)

## specialized the insignificant value according to the significant level
corrplot(cor(corr_data), p.mat = res1[[1]], insig="blank")

###plot the  correlation matrix for the scaled and transformed data #####################

# The final scaled and transformed data :
transformedandscaleddata<-read.csv("worldbank_transformed_csv.csv")
finaldata<-transformedandscaleddata
## Remoive HealthExpGDp as contains the same information as the transformed variable
#logHealthexppercapita
finaldata$Healthexp.GDP<-NULL

# Change the name of Australia to Oceania :
finaldata$Continent<-as.character(finaldata$Continent)
index<-which(finaldata$Continent=="Australia")
finaldata[index,length(finaldata)]<-"Oceania"
## Assign the countries to the rownames
finaldata1<- finaldata[,-1]
rownames(finaldata1) <- finaldata[,1]
finaldata<-finaldata1



# Change irrleelvant variables
finaldata$Rule.Type<-NULL
finaldata$Flag.Colours<-NULL
finaldata$Flag.Colours.Raw<-NULL
finaldata$Official.Languages<-NULL
finaldata$Official.Languages.Raw<-NULL
```

```r
data_nocat<-finaldata[1:19]
#str(finaldata)
#str(data_nocat)

## Look at correlation plots instead


corr_data1<-data_nocat
colnames(corr_data1)<-abbreviate(colnames(corr_data1))
#corrplot(cor(corr_data1), method="ellipse")


res2 <- cor.mtest(corr_data1,0.95)

## specialized the insignificant value according to the significant level
corrplot(cor(corr_data1), p.mat = res2[[1]], insig="blank")


## use the covariance matrix
data_nocat1<-data_nocat
rownames(data_nocat1)<-abbreviate(rownames(data_nocat1))
colnames(data_nocat1)<-abbreviate(colnames(data_nocat1))
pcawithout_scaling<-princomp(data_nocat1,cor=F)



par(mfrow=c(1,1))
## Consider PCa with scaling correlation matrix =T
## Still use the correlation matrix ?
pcawith_scaling<-princomp(data_nocat1,cor=T)



##Scree plots
pr.var <- (pcawith_scaling$sdev)^2
# Variance explained by each principal component: pve
pve <- pr.var /sum(pr.var)

##
##for the unscaled data:

pr.noscalevar<-(pcawithout_scaling$sdev)^2
pvenoscale<-pr.noscalevar/(sum(pr.noscalevar))


###Plot the pve and cummulative prinicipal experience:

# Plot variance explained for each principal component
plot(pve, xlab = "Principal Component",
     ylab = "Proportion of Variance Explained",
     ylim = c(0, 1), type = "b")
```

```
points(pvenoscale,col="red",type="b")


legend("topright",col=c("black","red"),pch=16,c("PCA-Correlation matrix","PCA- Covariance Matrix"),cex=


par(mfrow=c(1,2))
biplot(pcawith_scaling,main=" PCA -corr matrix",cex=0.7)
biplot(pcawithout_scaling,main="PCA - cov matrix",cex=0.7)


eigv<- eigen(cor(data_nocat))$values
eigv>mean(eigen(cor(data_nocat))$values)
require(scales)
par(mfrow=c(1,1))
pca_score1<-pcawith_scaling$scores[,1]

pca_score2<-pcawith_scaling$scores[,2]

plot(pca_score1,pca_score2,pch=16,col=cscale(finaldata$logGDPperCapCurrUS.,
                                             seq_gradient_pal("red", "green")))

legend("topright",col=c("red","green"),pch=16,c("Low GDP","High GDP"),cex=0.4)

text(pca_score1,pca_score2, labels=abbreviate(finaldata[,23]),cex=0.6)

plot(pca_score1,pca_score2,pch=16,col=cscale(finaldata$logCaptureFishprodM3tonspertotpopulation,seq_gra

legend("topright",col=c("red","green"),pch=16,c("Low Capture","High Capture"),cex=0.4)

text(pca_score1,pca_score2, labels=abbreviate(finaldata[,22]),cex=0.6)


###Code for the table of loadings
require(pander)
#install.packages("plyr")
#require(plyr)
pca_corr<-princomp(data_nocat,cor=T)

pca_loadings<-unclass(loadings(pca_corr))

Mardia1stpc<-which(abs(pca_loadings[,1])>0.7*max(abs(pca_loadings[,1])))


##
loadinsonly<-pca_loadings
LoadingsPC1<-loadinsonly[,1]
Mardiaspc1laodings<-LoadingsPC1[Mardia1stpc]


# Using Mardia's condition we obtain 9 variables which are important to the  1st principal component
#The variable "logHealthexpperCap","logGDPperCapCurrUS", and"InternetUsersper100"  serves to be the var
# all variables negative except for the "Adolesfertillity per100".
```

```r
# Similarly we can look at the 2nd compoenent using the Mardias condition L


## Look at the 2nd component for the loadings

Mardia2ndpc<-which(abs(pca_loadings[,2])>0.7*max(abs(pca_loadings[,2])))


LoadingsPC2<-loadinsonly[,2]
Mardiaspc2laodings<-LoadingsPC2[Mardia2ndpc]


## Most important variable in the 2nd PC : AgriculatureLand area
# Contrast beween Forestarea  and the Agriculture Land Area and  CO2kgper2010
#US$ of GDP  , a decrease in Agriculture Land Area and CO2 emissions that stem from
#burning of fossil fuels and the manufacture of cement  leads to an increase in the Forest
#Area of the countries .


## Look at the 3rd component for the loadings :

MardiasPC3<-which(abs(pca_loadings[,3])>0.7*max(abs(pca_loadings[,3])))

loadingsPC3<-loadinsonly[,3]
Mardiapc3loadings<-loadingsPC3[MardiasPC3]


## Look at the 4th component for the loadings
MardiasPC4<-which(abs(pca_loadings[,4])>0.7*max(abs(pca_loadings[,4])))

loadingsPC4<-loadinsonly[,4]
Mardiapc4loadings<-loadingsPC4[MardiasPC4]



## On the 4th Component , an increase in Population density , leads to a decrease in Forest Area


##LOokking at the 5th componnet :
MardiasPC5<-which(abs(pca_loadings[,5])>0.7*max(abs(pca_loadings[,5])))

loadingsPC5<-loadinsonly[,5]
Mardiapc5loadings<-loadingsPC5[MardiasPC5]




## Looking at the 6 th compoennet :
MardiaPC6<-which(abs(pca_loadings[,6])>0.7*max(abs(pca_loadings[,6])))
loadingsPC6<-loadinsonly[,6]
Mardiapc6loadings<-loadingsPC6[MardiaPC6]
```

```
a<-data.frame(names(pca_loadings[,1]))

b<-data.frame(Mardiaspc1laodings)
c<-data.frame(Mardiaspc2laodings)
d<-data.frame(Mardiapc3loadings)
e<-data.frame(Mardiapc4loadings)
f<-data.frame(Mardiapc5loadings)
g<-data.frame(Mardiapc6loadings)



set.caption("loadings of the first 6 components on Mardia's criterion")
pander(list(`PC1 Loadings on MArdia`=b,`PC2 Loadings on MArdia`=c,`PC3 Loadings on MArdia`=d,`PC4 Loadi
```

References

Gentleman, Robert, Kurt Hornik, Giovanni Parmigiani, and H Wickham. 2008. *Use R !* doi:10.1007/978-0-387-78171-6.

Lai, Dejian. 2003. "Principal component analysis on human development indicators of China." *Social Indicator Research* 61: 319–30.

Shah, Anoop D., Jonathan W. Bartlett, James Carpenter, Owen Nicholas, and Harry Hemingway. 2014. "Comparison of random forest and parametric imputation models for imputing missing data using MICE: A CALIBER study." *American Journal of Epidemiology* 179 (6): 764–74. doi:10.1093/aje/kwt312.