Multivariate Methods Coursework

PhaniKanchan Adabala

Introduction:

This coursework aims to analyse the data of various health, economic, and environmental indicators of different countries in the world. Our objectives are to:

a) Make the data analysable in R

b) Analysing the data with multivariate methods by discovering the dependency between the variables and groups of variables.

c) Performing the dimension reduction with Principal component Analysis

d) Perform canonical correlation to find how many dimensions are necessary to understand the correlation between two sets of variables i.e., across the health, economic and environmental indicators.

The data for health, economic and environmental indicators are collected for different years ranging over a period of time from 1995 to 2021. The data collected in the respective year is indicated in the brackets below. A total of 4 health, 7 economic and 3 environmental indicators are chosen.

Health indicators include :

a) **NB_mortality**: New-born Mortality rate, children dying within the first 28 days per 1000 live births (2015).

b) **BABY_perwom**: Babies per woman indicating the total fertility rate (2013).

c) **Vaccined_percent**: DTP3 immunized i.e., the percentage of the 1-year-old children who received 3 vaccine doses (2019).

d) **GovtHealth_spend** : Govt. Health spending per person is expressed in US dollars (2009).

Economic indicators :

a) **GNI_percapita**: Gross national income per capita based on purchasing power parity (2015).

b) **child_elder**: Dependency Ratio i.e., the ratio of children and elderly population to normal aged population (2021).

c) **CPI** : Corruption perception Index (2017) .

d) **GDP_percapita**: Average growth in GDP Per capita over the subsequent 10 years (1995).

e) **Exports_percent**: Exports expressed in terms of percentage of GDP (2018).

f) **FI_inflows**: Foreign Investment Inflows expressed as a net percentage of GDP (2014).

g) **Female_15.** : contribution of females aged 15+ to the labour force (2017) 2019.

Environmental Indicators :

a) **cell_phones**: Cell phones per 100 people indicating the total active subscriptions to cellular telephones (2015).

b) **Electric**: Electricity per person shows the usage of electricity per person (2014).

c) **Co2_emission**: Co2 emissions expressed in 1000 metric tonnes (2018).

Additionally, the data for the grouping of countries is also collected. The countries can be grouped according to their geographical regions such as East Asia and Pacific (EAPAC), Europe and Central Asia (ECA), Latin America and the Caribbean (LAC), Middle East and North Africa (MENA), North America (NOA),– South Asia (SA), Sub-Saharan Africa (SSA). Or can be grouped according to their income which is Low income (LOW), Lower-middle income (LOM), Upper-middle income (UPM), High income (HIGH). These features are captured as categorical variables "Geog" and "Income".

Data Quality & characterisation:

The data of all the indicators for the respective years are combined to make a single file. There were some missing values in each indicator for which the equivalent rows are removed so that data for every country is there for every indicator. Also, the data is made compatible such that all the notations are converted into a number format. The data also contained special characters which were removed. The total number of countries that were obtained after completing all the data cleaning is 125.

Analysis:

We can find the dependency between the variables by plotting a scatter plot. Scatter plots show that the newborn mortality rate is strongly, and positively correlated with babies born per woman. A correlation matrix of the variables provides the magnitude of the correlation. we can conclude that the variables that have a strong correlation are BABY_perwom with child_elder (0.8), GNI_percapita with GovtHealth_spend (0.75) and with Co2_emission (0.75). The variables that show less correlation with others are FI_inflows, and GDP_percapita.

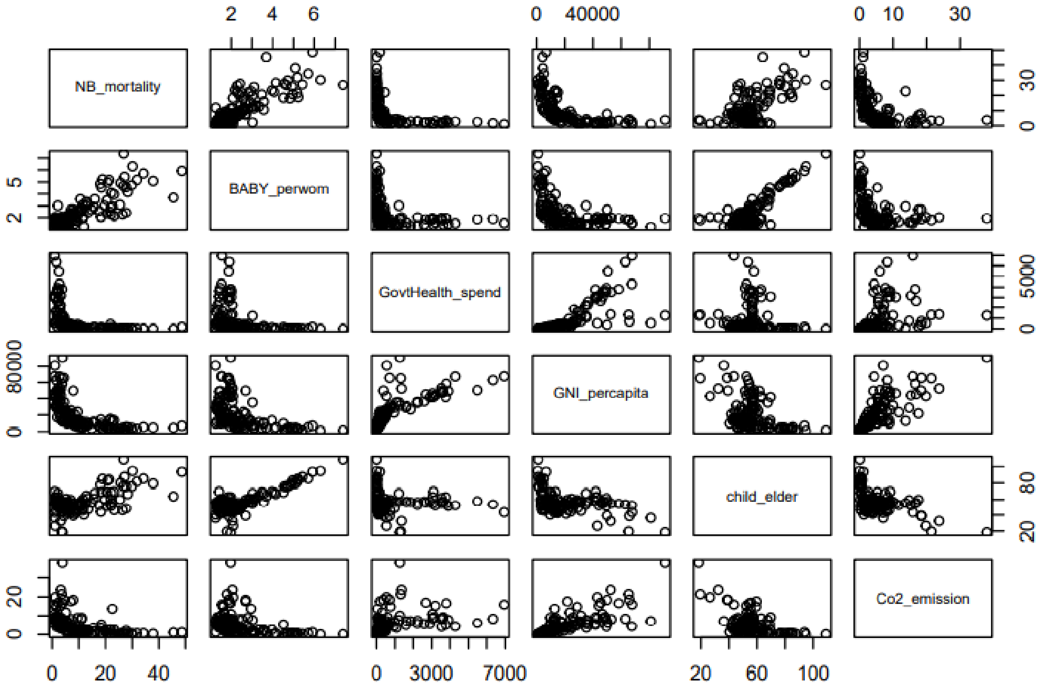

Figure-1: Scatterplots of the highly correlated variables

Since the variables are highly correlated among themselves, there is a need to reduce the dimension of the data in order to build a model or do any analysis. The principal component analysis is the best method to use to reduce the dimension, where there are many variables in the data and have high correlations among them. Since the data is not commensurate, we first standardise the data and then perform the analysis.

The eigen values of the covariance matrix represent the principal components. The first eigen value is very large and represents the maximum variance of the data. The corresponding eigen vector of this eigen value is the vector of linear combination of variables.

```
## [1] 5.78127299 1.85963772 1.19313360 0.99468211 0.85339141 0.77616507

## [7] 0.63506372 0.55528491 0.44997982 0.35925173 0.23566736 0.17498952

## [13] 0.07580927 0.05567077
```

The maximum variation of the data is present along the PC1 axis, and less on PC2 and PC3. The percentage contribution of each principal component is shown in figure 2.

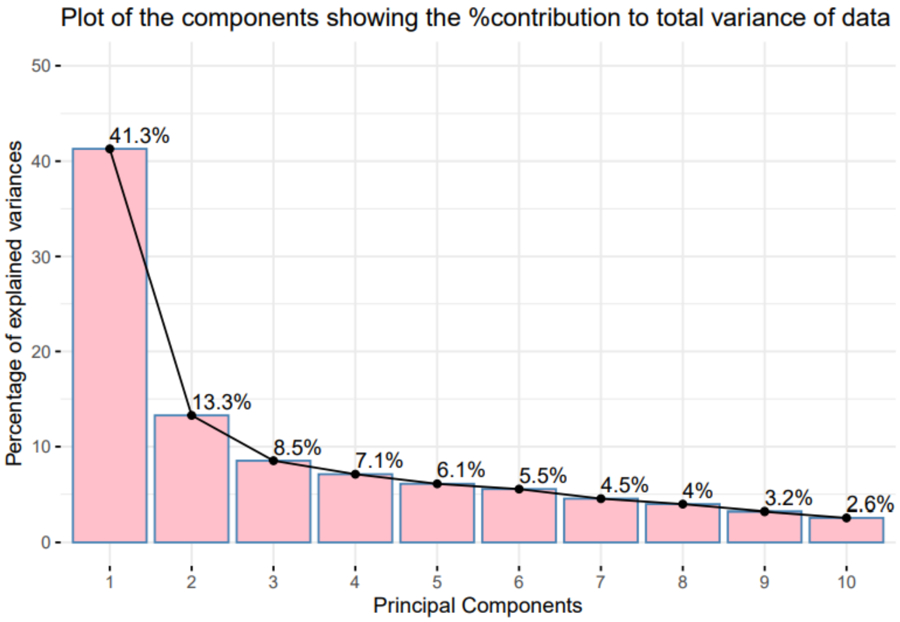

Figure-2: Percentage contribution to variance by principal components

The first principal component contributes to 41.3% of the total variance of the data and the second contributes to 13.3%, the third to about 8.5% and so forth. But how many dimensions should be considered? usually, the eigenvalues that are greater than the mean of all the eigenvalues should be considered to explain the good amount of variance in the data. By this rule, we should consider all the eigenvalues that are greater than 1 which means the first three principal components. Then the total % of variance captured by these three principal components is 63.1% which is less than 70% and the fourth eigenvalue is much closer to 1 (0.995), so also considering this would lead to about 70.2% of the total variance.

Also, we can identify which variables contribute to how much in each principal component. The figures below represent the contribution of each variable in Principal component 1 and principal component 2. The red dashed lines in Figures 3 and 4 explain the expected average contribution of the variables.

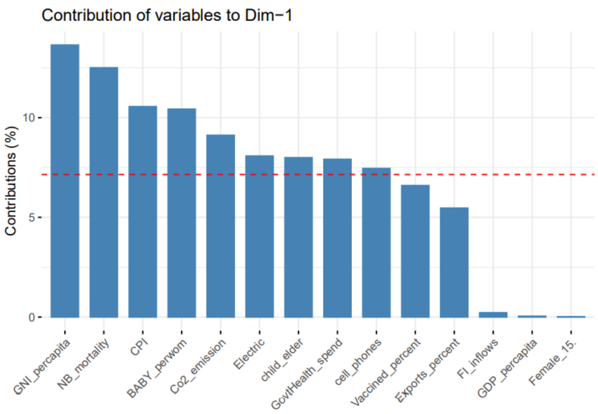

Figure-3: % contribution of variables to PC1

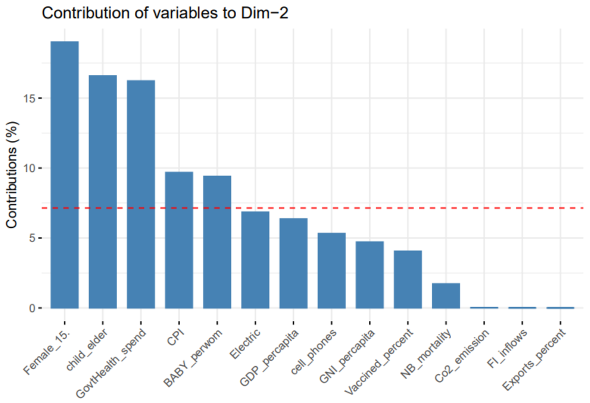

Figure-4: % contribution of variables to PC2

It can be seen that the Economic indicators GNI_percapita, CPI and health indicators NB_mortality, Baby_perwoman contribute to the Principal component 1. While in principal component 2, the Environmental indicator Female 15+ shows the highest contribution followed by the child_elder ratio which is the Economic indicator and health indicator of Govt. Health spending. In both the principal components, FI_inflows has a very less contribution.

Now, the correlations between the variables and the principal components can be visualised with the help of a variable correlation plot. Also, the contribution of the variables can easily be visualised. Figure 5 represents the variable correlation plot of the data.

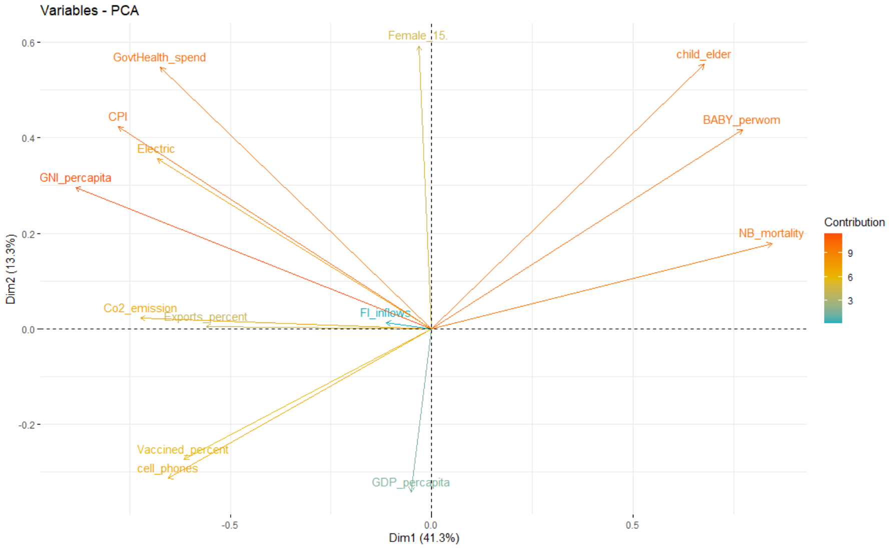

Figure-5: Variable contribution plot

The most contributing variables are the ones marked in red colour i.e., GNI_percapita, BABY_perwom, CPI and NB_mortality whereas the least contribution is from FI_inflows, GDP_percapita, Exports_percent etc., The variables in yellow colour show slight contribution.

Now, visualising the scores of the observations in the first two principal components. when the plot is made dimension 1 against dimension 2, we get some nice butterfly-like structure explaining the maximum variance along with the first principal component and 13.3% along with the second principal component. The other two dimensions sum up to about 15.6% which can be quite hard to show visually. The plotted scores are grouped according to their geographic regions as well as by their Income.

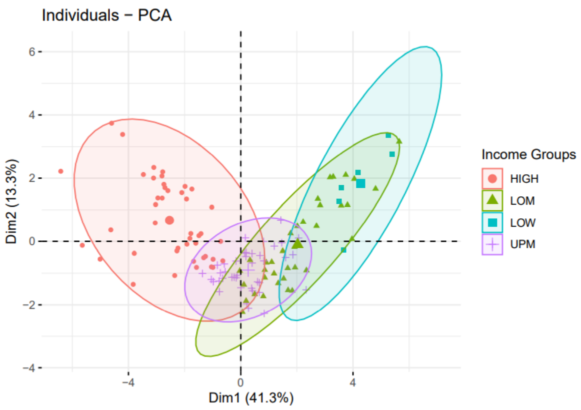

Figure-6: Scores plotted and grouped by Income

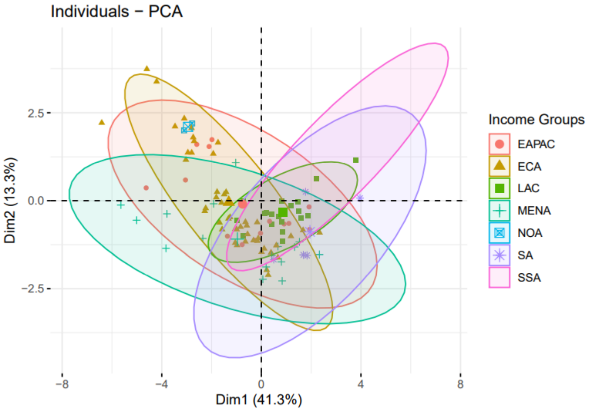

Figure-7: Scores grouped Geographically

The variables in the variable correlation plot and the observation scores can be grouped into a single plot known as a biplot. The biplot for the data division across the Income groups is shown in figure 8.

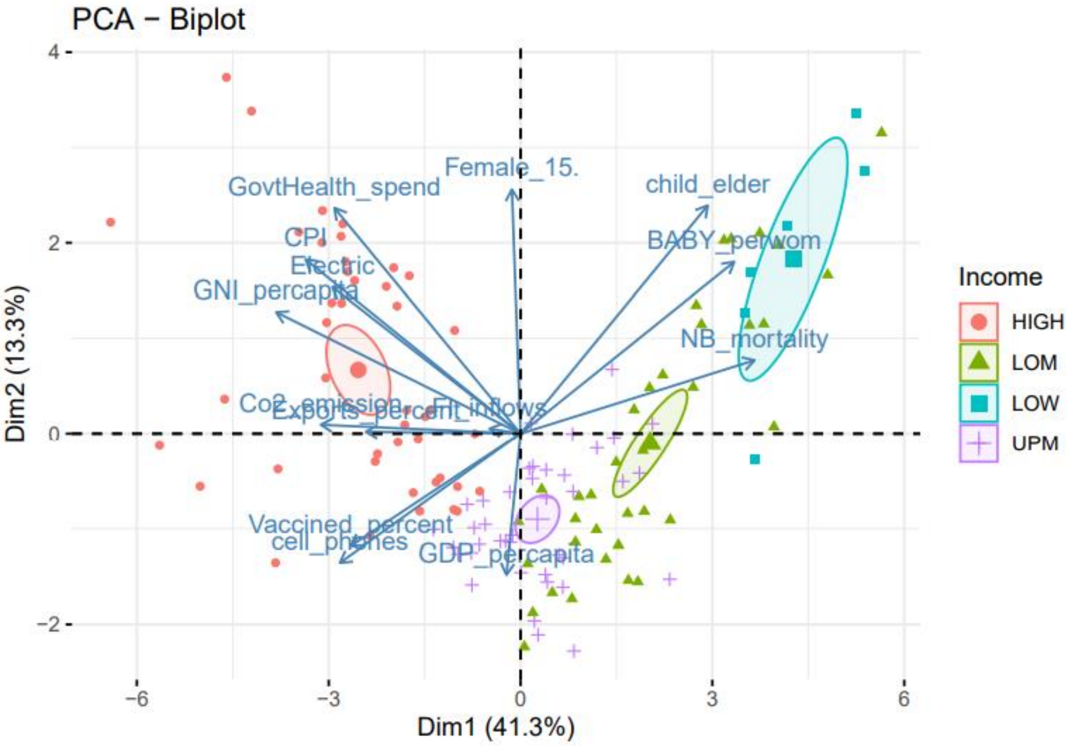

Figure-8: Biplot of the total data

From the above biplot, we can infer that the observations formed natural groups by their income-wise. The short ellipses represent the group means for the income groups. The plotted values are the scores of the observations. The vectors from the origin represent the variables in the data. The ones that are far from the origin have a strong contribution, while the ones near to the origin are less important. The variables that are grouped in the same quadrant are positively correlated with each other and the ones negatively correlated are in the opposite quadrant. The variables that are at right angles to each other are uncorrelated.

Positively correlated variables :

1) Child_elder, BABY_perwom, NB_mortality.
2) Vaccined_percent, cell_phones, GDP_percapita
3) GovtHealth_spend, Female_15. , CPI, Electric, GNI_percapita, Co2_emission, Exports_percent, FI_inflows.

Negatively correlated variables :

Child_elder, BABY_perwom, NB_mortality with Vaccined_percent, cell_phones, GDP_percapita

Uncorrelated variables :

1) GovtHealth_spend, child_elder
2) GovtHealth_spend, cell_phones

Now, the canonical correlation can be done to understand what variables are necessary to define the maximum correlation between two sets of variables. First, considering the Health and Economic variables, we measure the linear relationship between these two groups of multidimensional variables.

The coefficient vectors of the canonical correlation between health and economic indicators are :

CV_health = - 0.707 * NB_mortality + 0.0606* BABY_perwom - 0.0876* Vaccined_percent + 0.698* GovtHealth_spend

CV_Economic = - 0.727* GNI_percapita + 0.674* child_elder + 0.055* CPI + 0.111* GDP_percapita- 0.016* Exports_percent - 0.034* FI_inflows + 0.012* Female_15.

The variables that contribute to the maximum correlation between two sets of variables are the Newborn Mortality Rate and Government health spending from health indicators, GNI_percapita and Child_elder dependency ratio from Economic indicators. The maximum correlation is given by the square root of the largest eigenvalue which is 0.94.

The corresponding canonical variates are calculated and plotted against each other as shown in figure 9.

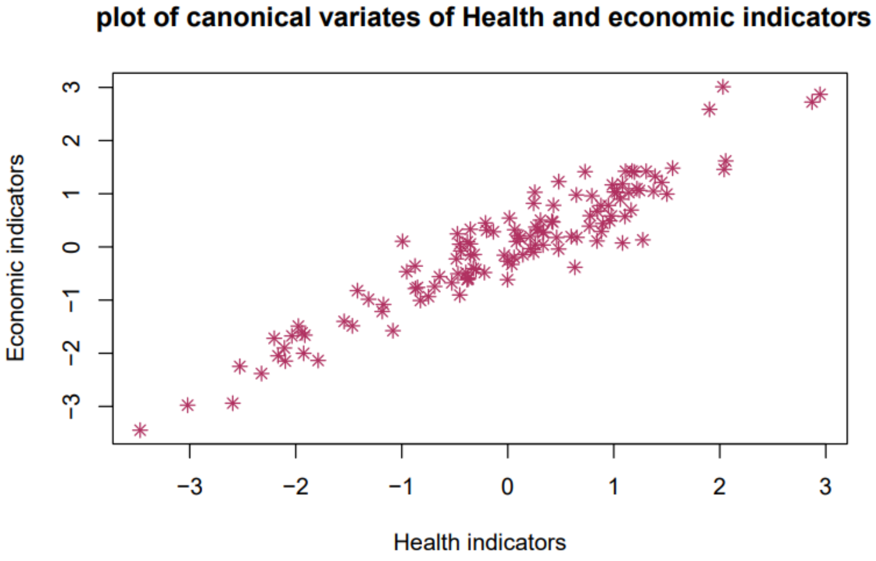

Figure-9: Canonical variates of Health and Economic indicators

Conclusion:

The dependency between the variables is clearly articulated. The data containing 14 variables was successfully reduced to 4 dimensions which are the linear combinations of the variables. The new dimensions of the data can be interpreted as below :

PC1 = 0.354* NB_mortality + 0.323* BABY_perwom - 0.257* Vaccined_percent -

0.281* GovtHealth_spend - 0.369* GNI_percapita + 0.283* child_elder - 0.325* CPI -

0.234* Exports_percent - 0.273* cell_phones - 0.284* Electric - 0.302* Co2_emission

PC2 = 0.131* NB_mortality + 0.307* BABY_perwom - 0.201* Vaccined_percent +

0.403* GovtHealth_spend + 0.217* GNI_percapita + 0.407* child_elder + 0.311* CPI –

0.252* GDP_percapita + 0.436* Female_15 - 0.231* cell_phones + 0.262* Electric

PC3 = -0.382*GDP_percapita - 0.480*Exports_percent - 0.681*FI_inflows - 0.315*Female_15. –

0.161 * Electric + 0.148* Co2_emission

PC4 = -0.136*Vaccined_percent + 0.106*child_elder + 0.796*GDP_percapita +

0.122*Exports_percent - 0.425*FI_inflows -0.197*cell_phones + 0.199 * Electric +

0.196 * Co2_emission

Among which the major variation lies in PC1. The data is now transformed into four variables that represent about 70.2% of the total variance. The canonical correlation between Health and Economic indicators is maximised to a value of 0.94 and the variables responsible for this are NB_mortality, GovtHealth_spend, GNI_percapita and child_elder.

Appendix

```
##################################
############# Correlations #######
##################################

options(warn=-1)
df <- read.csv("Coursework.csv",header = TRUE)
pairs(df[c(2,3,5,6,7,15)])

##################################
################ PCA #############
##################################

library(dplyr)
library(factoextra) #fviz_contrib
library(FactoMineR) #fviz_pca_biplot
library(ggfortify)  #autoplot

dfn <- df[2:15] %>% mutate_all(~(scale(.) %>% as.vector))
dfn_mat <- as.matrix(dfn)
y_bar <- apply(dfn_mat,2,mean)
S <- cov(dfn_mat)

eig_val <- eigen(S)$values
eig_vec <- eigen(S)$vectors

PCA <- princomp(dfn_mat)
PCA$loadings

##################################
######## PCA Visualisations ######
##################################

# eigen values plot
fviz_eig(PCA,barfill = "pink", addlabels = TRUE, ylim = c(0, 50),
         xlab = "Principal Components",
         title  = "Plot of the components showing the %contribution
         to total variance of data")

# contribution of variables to PC1
fviz_contrib(PCA, choice = "var", axes = 1, top = 15)

# contribution of variables to PC2
fviz_contrib(PCA, choice = "var", axes = 2, top = 15)

# observations plot
fviz_pca_ind(PCA,
             geom.ind = "point",
             col.ind = df$Income, # color by Income
             addEllipses = TRUE,
             legend.title = "Income Groups")

# VAriables plot
```

```
fviz_pca_var(PCA, col.var = "contrib",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             legend.title = "Contribution")

#Biplot
fviz_pca_biplot(PCA, geom.ind = "point",
                col.ind = df$Income,
                addEllipses = TRUE,
                ellipse.type = "confidence",
                legend.title = "Income")

###################################
##### Canonical Correlation ######
###################################

df_cr <- df[2:15] %>% mutate_all(~(scale(.) %>% as.vector))
cor(df_cr)

S <- cov(df_cr)
Syy <- S[2:5,2:5]
Sxx <- S[6:12,6:12]
Syx <- S[2:5,6:12]
Sxy <- S[6:12,2:5]

Sy <- solve(Syy) %*% Syx %*% solve(Sxx) %*% Sxy
Sy_EVal <- eigen(Sy)$values
Sy_EVec <- eigen(Sy)$vectors

sqrt(max(Sy_EVal))
# this is the pearson correlation coefficient for the two sets of
# variables i.e, health and economic indicators

# The eigen vector of the largest eigen value is the vector a,
# which is used to express the variables as a linear combination
eigen(Sy)$vectors[,1]

# calculate the vector a for the economic variables
Sx <- solve(Sxx) %*% Sxy %*% solve(Syy) %*% Syx
Sx_EVal <- eigen(Sx)$values
Sx_EVec <- eigen(Sx)$vectors

# The standardized vector of Economic indicators
eigen(Sx)$vectors[,1]

CV_health <- as.matrix(df_cr[2:5]) %*% as.matrix((eigen(Sy)$vectors[,1]))
CV_Econom  <- as.matrix(df_cr[6:12]) %*% as.matrix((eigen(Sx)$vectors[,1]))
plot(CV_health ,CV_Econom,
     xlab = "Health indicators",
     ylab = "Economic indicators",
     main = " plot of canonical variates of Health and economic indicators",
     pch = 8,col = "Maroon")
```

References

1) Based on free material from GAPMINDER.ORG, CC-BY LICENSE.[Online].[Accessed 26 April 2022].Available from: https://www.gapminder.org/data/

2) [Online].[Accessed 26 April 2022].Available from: https://datahelpdesk.worldbank.org/knowledgebase /articles/906519-world-bank-country-and-lending-groups

3) [Online].[Accessed 26 April 2022].Available from: http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/112-pca-principal-component-analysis-essentials/