ANALYSIS OF COUNTRIES USING MULTIVARIATE METHODS

1 INTRODUCTION

In today's world, tons of indices and variables are used by analysts to determine the position of countries in the scheme of things and also to get a feeling of what takes place in each country. The aim of this report is to analyse fourteen (14) indices in countries around the world, grouped into three (3) categories which are, health, economic, and environmental. These variables will be analysed with the aid of multivariate methods like canonical correlation, principal component analysis and discriminant analysis. The data used for this analysis was obtained from the Gapminder Foundation (http://www.gapminder.org/data/).

One hundred and eighteen (118) countries (observations) will be analyzed based on 14 variables. These variables and the years they were acquired are as follows:

- **Health variables**
  - Newborn mortality rate per 1000 births (newbornMortality): Newborn mortality rate in all countries from the year 2015.
  - Babies per woman (childPerWoman): Babies per woman in all countries from the year 2022.
  - Diphtheria tetanus toxoid and pertussis (DTP3) immunized (dtp3Immunized): Percentage of one year olds that are DTP3 immunized (World Health Organization, 2022). The data was obtained from the year 2019.
  - Government Health spending per person in US dollars ($) (healthSpending) from the year 2009.
- **Economic variables**
  - Gross National Income (GNI), Purchasing Power Parity, per capita, in US $ (gniPPP) from the year 2019.
  - Proportion of children and elderly per 100 adults (childElderPer100) from the year 2022.
  - Corruption Perception Index (cpi) : Perceived levels of corruption within a country with low values indicating high corruption and higher values indicating the opposite (Transparency International, 2021). The cpi data was obtained from 2017.
  - Gross Domestic Product (GDP) per capita growth of the next ten (10) years (gdp10years) from the year 2001.
  - Exports(percentage of GDP) for every country (perExportsGDP) from the year 2019.
  - Foreign investment inflows (percentage of GDP) for every country (foreignInvest) from the year 2019.
  - Females aged 15 and higher participating in the labour force (femaleAge15) from the year 2022.
- **Environmental variables**
  - Number of cell phones per 100 people in each country (phonePer100) from the year 2019.
  - Electricity or power use per person in each country (electricUse) from the year 2014.
  - Carbon(iv)oxide(CO_2) emissions in tonnes per person in each country (co2Emissions) from the year 2018.

Also, the countries are grouped into two different categories, Geography and Income. In the Geography category, thirteen (13) countries are in East Asia and the Pacific (EAPAC), forty-five (45) countries in Europe and Central Asia (ECA), nineteen (19) countries in Latin America and the Caribbean (LAC), sixteen (16) countries in the Middle East and North Africa (MENA), two (2) countries in North America (NOA), four (4) countries in South Asia (SA) and 19 countries in Sub-Saharan Africa (SSA). In the case of income, forty-four (44) countries are High income (HIGH), thirty-eight (38) countries are Upper-Middle income (UPM), thirty-two (32) countries are Lower-Middle income (LOM), and 4 countries are Low income (LOW).

The countries, with the fourteen variables, will be further explored and analysed in subsequent sections.

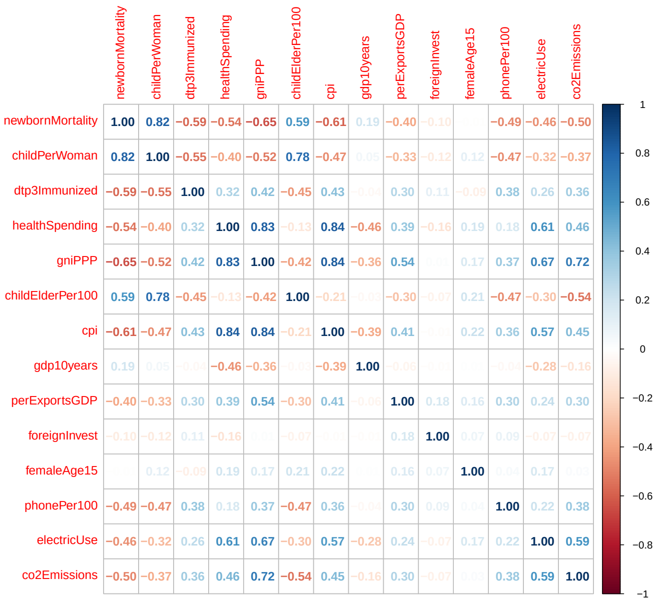
2 ANALYSIS AND RESULTS

In this section, the variables will the explored and possible dependencies or correlations will be unearthed between them. Table 1 show the descriptive statistics of all 14 variables.

Table 1: Descriptive Statistics of all variables with number of observations (n) = 118.

S/N	Variable	Unit	Mean	Variance	Standard Deviation
1	newbornMortality	per 1000 births	10.32	99.95	10.00
2	childPerWoman	Babies per woman	2.26	0.95	0.97
3	dtp3Immunized	Percent of 1 year olds	90.33	99.16	9.96
4	healthSpending	US $	938.22	1,342,108	1,158.49
5	gniPPP	US $	26,586.02	472,516,132	21,737.44
6	childElderPer100	per 100 people	55.68	163.97	12.81
7	cpi	None	46.91	354.34	18.82
8	gdp10years	percentage growth	3.01	5.26	2.29
9	perExportsGDP	percent of GDP	44.59	911.54	30.19
10	foreignInvest	percent of GDP	4.16	141.51	11.90
11	femaleAge15	work force percentage	50.71	190.59	13.81
12	phonePer100	per 100 people	118.91	820.29	28.64
13	electricUse	per person	4637.94	40,075,616	6,330.53
14	co2Emissions	tonnes per person	5.69	35.42	5.95

Following this, the correlation between variables will give insight into how each variable relates to one another. From figure 1, countries with high newborn mortality rates tend to have high fertility rates. On the other hand, countries with a high GNI per capita tend to spend more on health, have high corruption indexes, use up a lot of electricity and emit a lot of CO_2. Also, variables foreignInvest and femaleAge15 have little or no correlations with other variables.



	newbornMortality	childPerWoman	dtp3Immunized	healthSpending	gniPPP	childElderPer100	cpi	gdp10years	perExportsGDP	foreignInvest	femaleAge15	phonePer100	electricUse	co2Emissions
newbornMortality	1.00	0.82	−0.59	−0.54	−0.65	0.59	−0.61	0.19	−0.40	−0.10		−0.49	−0.46	−0.50
childPerWoman	0.82	1.00	−0.55	−0.40	−0.52	0.78	−0.47	0.05	−0.33	−0.12	0.12	−0.47	−0.32	−0.37
dtp3Immunized	−0.59	−0.55	1.00	0.32	0.42	−0.45	0.43	−0.04	0.30	0.11	−0.09	0.38	0.26	0.36
healthSpending	−0.54	−0.40	0.32	1.00	0.83	−0.13	0.84	−0.46	0.39	−0.16	0.19	0.18	0.61	0.46
gniPPP	−0.65	−0.52	0.42	0.83	1.00	−0.42	0.84	−0.36	0.54	0.01	0.17	0.37	0.67	0.72
childElderPer100	0.59	0.78	−0.45	−0.13	−0.42	1.00	−0.21	−0.03	−0.30	−0.07	0.21	−0.47	−0.30	−0.54
cpi	−0.61	−0.47	0.43	0.84	0.84	−0.21	1.00	−0.39	0.41	−0.01	0.22	0.36	0.57	0.45
gdp10years	0.19	0.05	−0.04	−0.46	−0.36	−0.03	−0.39	1.00	−0.06	−0.01		−0.04	−0.28	−0.16
perExportsGDP	−0.40	−0.33	0.30	0.39	0.54	−0.30	0.41	−0.06	1.00	0.18	0.16	0.30	0.24	0.30
foreignInvest	−0.10	−0.12	0.11	−0.16	0.01	−0.07	−0.01	−0.01	0.18	1.00	0.07	0.09	−0.07	−0.07
femaleAge15		0.12	−0.09	0.19	0.17	0.21	0.22		0.16	0.07	1.00	0.04	0.17	0.03
phonePer100	−0.49	−0.47	0.38	0.18	0.37	−0.47	0.36	−0.04	0.30	0.09	0.04	1.00	0.22	0.38
electricUse	−0.46	−0.32	0.26	0.61	0.67	−0.30	0.57	−0.28	0.24	−0.07	0.17	0.22	1.00	0.59
co2Emissions	−0.50	−0.37	0.36	0.46	0.72	−0.54	0.45	−0.16	0.30	−0.07	0.03	0.38	0.59	1.00

Figure 1: Correlation matrix of all variables.

More insight can be drawn by finding dependencies between groups of variables. In the previous section, it was stated that all variables are divided into three (3) groups (Health, Economic, and Environmental).

The relationships will be unearthed with the aid of canonical correlation.

2.1 CANONICAL CORRELATION ANALYSIS OF THE VARIABLES

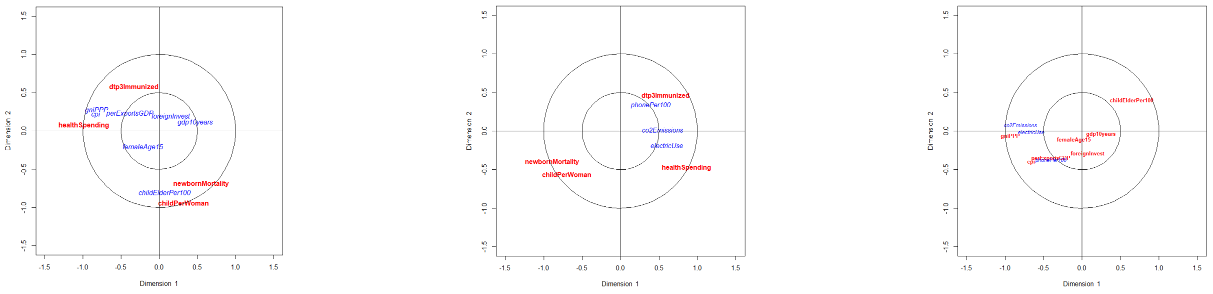
Canonical correlation helps to determine if groups of variables are correlated to one another. Although, before canonical correlation can occur, an assumption of normality has to be made. All variables in the data set being analysed will be assumed to have a multivariate normal distribution based on the central limit theorem as the number of observations (118) is deemed to be large enough. Below is a summary of the canonical correlation between each of the three groups of variables and the statistical significance of each correlation at the 5% significance level. The test being used is known as the Wilks Lambda test.

Table 2: Canonical Correlations between groups of variables

HEALTH VS ECONOMIC VARIABLES			
Dimension	Canonical Correlation	P-value	Statistically Significant(Yes/No)
1	0.911	0.000	Yes
2	0.852	0.000	Yes
3	0.188	0.757	No
4	0.154	0.614	No
HEALTH VS ENVIRONMENTAL VARIABLES			
Dimension	Canonical Correlation	P-value	Statistically Significant(Yes/No)
1	0.682	0.000	Yes
2	0.434	0.000	Yes
3	0.090	0.630	No
ECONOMIC VS ENVIRONMENTAL VARIABLES			
Dimension	Canonical Correlation	P-value	Statistically Significant(Yes/No)
1	0.841	0.000	Yes
2	0.452	0.000	Yes
3	0.362	0.007	Yes

From table 2, in the first two groups of variables, only the first two dimensions of their respective canonical correlation values are statistically significant. This means there exists a correlation between the two sets of variables in each of those groups based on the first two dimensions. The same can be said for the last set (Economic and Environmental) as all its canonical correlation dimensions are statistically significant.

Figures 2a, 2b, and 2c below show the plots, obtained by the "CCA" package on R, of the direction of the variables on the first two (2) canonical correlation dimensions for each set of variable groups. The plots give insight into how each variable affects the canonical correlation for each dimension. From figure 2a, based on their canonical coefficients, variables gniPPP, healthSpending, cpi, and newbornMortality contribute to the first canonical covariate (dimension 1) of the Health and Economic variables. In figure 2b, variables newbornMortality, child_fertility and healthSpending contribute to the first canonical covariate value of the Health and Environmental variables. Finally, in figure 2c, gniPPP, co2Emissions, and electricUse heavily influence the first canonical covariate of the economic and environmental variables.



(a) Health and Economic **(b)** Health and Environmental **(c)** Economic and Environmental

Figure 2: Canonical Correlation plots between Groups of Variables

With the dependencies between the variables and groups of variables shown, the next step is to determine the amount of variability in the data set to uncover new information from the observations. This is achieved by reducing the dimensions of the data set through principal component analysis.

2.2 PRINCIPAL COMPONENT ANALYSIS OF THE DATA

Principal component analysis (PCA) aids in reducing the dimensions of data. Once the dimensions of the data set have been reduced, the variations in the data set can be observed and insights can be gained from the data. PCA will be applied to all 118 observations and 14 variables. However, the standardized version of PCA will be applied as, from table 1, the units of all 14 variables are not commensurate and thus, this will give a better understanding of the variations in the data.

The R package "ggbiplot" was used for the PCA of the data. Once PCA was applied to the data, fourteen (14) components were obtained (Principal Component (PC) 1 to PC14). The scree plot in figure 3 shows how much variation each PC has.

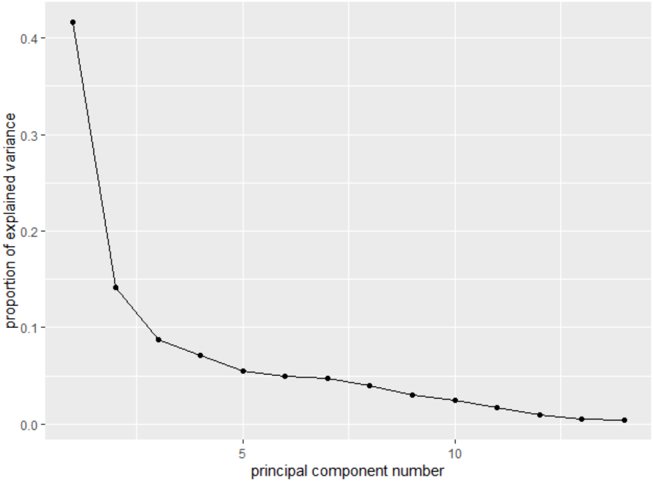


Figure 3: Scree plot of 14 principal components

From figure 3, PC1 and PC2 contain almost 60% of the variation in the data. These two component contain the most significant amount of information from the data. Due to this, PC1 and PC2 will be used to explain the variations in the data. To view this firsthand, the data will be plotted on axes containing both components. Figure 4 gives the plot of the data on both PC1 and PC2 as the x and y axes, along with the effect of the each variable (loading) represented as vectors on the graph.

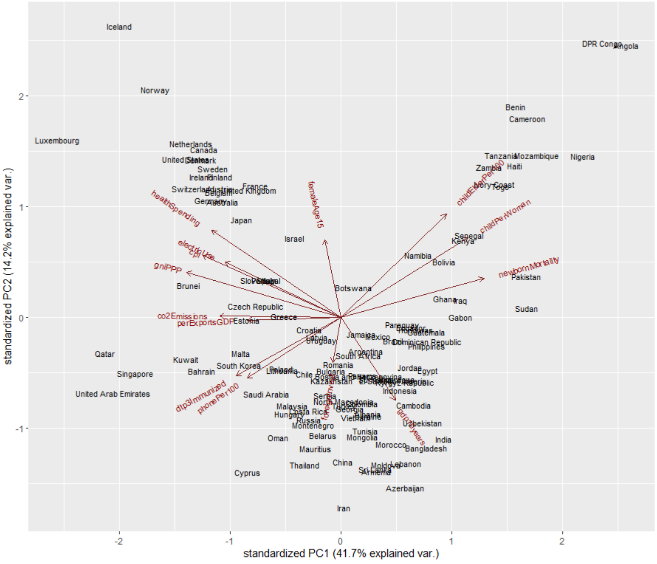


Figure 4: PC1 and PC2 plot of the Countries.

In figure 4, the data forms a nice ”V” shape after being plotted on both PC1 and PC2. Countries on the right side of the ”V” pattern are characterized by having high values for newbornMortality, childPerWoman, and childElderPer100. Conversely, countries on the left side of the ”V” pattern are characterized by high values for healthSpending, gniPPP, electricUse, and so on. Finally, countries at the base of the ”V” pattern tend to have high values for foreignInvest and gdp10years. Although, countries like Iceland, Iran, Luxembourg, and so on, are outliers in this PCA.

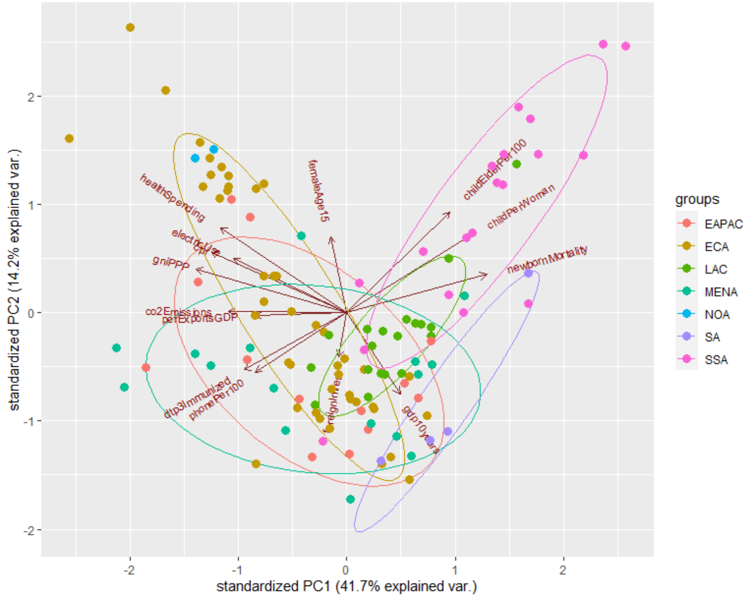


Figure 5: PC1 and PC2 plot of the Countries in Geographic Groupings

Further insight can be gained by looking at the countries in terms of their groups (Geography and Income). In figure 5, when the countries are placed in their respective geographic groupings, the countries in both ECA and NOA groups are characterized by their high values for healthSpending and gniPPP while on the other spectrum, countries in Sub-Saharan Africa (SSA) are characterized by high values for newbornMortality and child_fertility_woman.

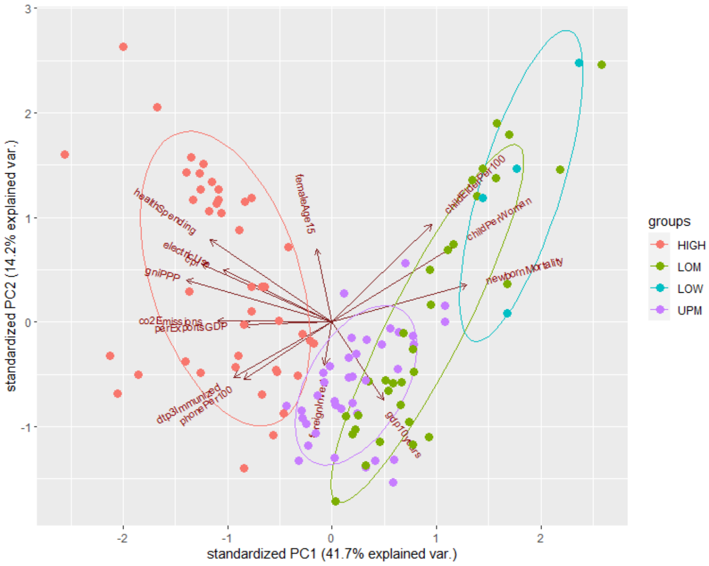


Figure 6: PC1 and PC2 plot of the Countries in Income Groupings

In figure 6, when countries are grouped by their incomes, Countries in the HIGH income group have similar features to those in ECA and NOA geographic groups, while countries in the LOW and LOM groups have similar features to those in SSA. Countries in the UPM income group are characterized with having high values for their gdp10years variable. On a final note, variable femaleAge15 seems to give no significant insight into countries in both principal components.

The final part of this section, will go over all patterns uncovered in the data and perform suitable methods to confirm the patterns.

2.3 PATTERNS IN THE DATA

From the previous analyses performed, patterns have emerged; some of which are both obvious and interesting. First, the canonical correlation analysis shows distinct relationships with the variables newbornMortality, childElderPer100, and childPerWoman. This means that countries with high values in one of the aforementioned variables have high values in the others. The same pattern exists with the variables gniPPP, healthSpending, cpi, electricUse, and co2Emissions.

Secondly, from the principal component analysis, a few patterns have been revealed from plotting the countries on the PC1 and PC2 axes. Based on the geography groupings in figure 5, countries with high values for gniPPP, healthSpending, and cpi are most likely to be in the ECA group. Countries with high values for newbornMortality, childElderPer100, and childPerWoman are likely to be located in SSA. It is worth noting that there is some clear overlap of countries in different groups.

With regards to Income, countries in the HIGH income group tend to have high values for gniPPP, healthSpending, cpi, and electricUse; as these countries tend to be more developed than others. Countries in the LOW and some in the LOM groups are characterized by having high newbornMortality, childElderPer100, and childPerWoman. Finally, countries in the UPM group have unusually high values for their gdp10years variable. This is probably due to the fact that countries in the middle class have the most potential to grow their GDP.

From the above patterns, it is heavily implied, with regards to the Income groupings, that countries in the HIGH and LOW incomes groups are disparate. The same can also be said for countries in the HIGH and LOM income groups. These can be confirmed with the aid of discriminant analysis. Discriminant analysis, implemented with the aid of R studio, is performed to confirm if the three (3) income groups can be separated.

Table 3: Discriminant Functions for the Separation of the HIGH, LOW and LOM income groups

S/N	Variable Name	Discriminant Function 1	Discriminant Function 2
1	newbornMortality	3.16×10^{-2}	-7.56×10^{-2}
2	childPerWoman	9.86×10^{-1}	1.06×10^{0}
3	dtp3Immunized	-3.72×10^{-3}	2.50×10^{-2}
4	healthSpending	3.07×10^{-5}	3.34×10^{-4}
5	gniPPP	-4.33×10^{-5}	3.79×10^{-5}
6	childElderPer100	-6.36×10^{-2}	2.62×10^{-2}
7	cpi	-3.73×10^{-2}	-2.24×10^{-2}
8	gdp10years	9.11×10^{-2}	5.52×10^{-2}
9	perExportsGDP	3.40×10^{-5}	2.03×10^{-3}
10	foreignInvest	-1.07×10^{-2}	1.96×10^{-2}
11	femaleAge15	-5.75×10^{-3}	-1.37×10^{-2}
12	phonePer100	3.25×10^{-3}	-2.95×10^{-2}
13	electricUse	1.08×10^{-6}	-2.52×10^{-6}
14	co2Emissions	-1.10×10^{-1}	2.17×10^{-2}

From table 3, the three groups are separated by two standardized discriminant functions (LD1 and LD2). In LD1, variables childPerWoman, newbornMortality, co2Emissions and gdp10years contribute to the separation of the three income groups, while in LD2, the same group of variables contribute to the separation in this function. Figure 9, displays the plots of LD1 and LD2 that try to separate the three groups. The plot discriminates the HIGH from the LOW and LOM income countries. This is mostly due to LD1. However, the discriminant functions fail to separate the LOW and LOM income countries. These findings are in-line with the findings acquired from the principal component analysis.

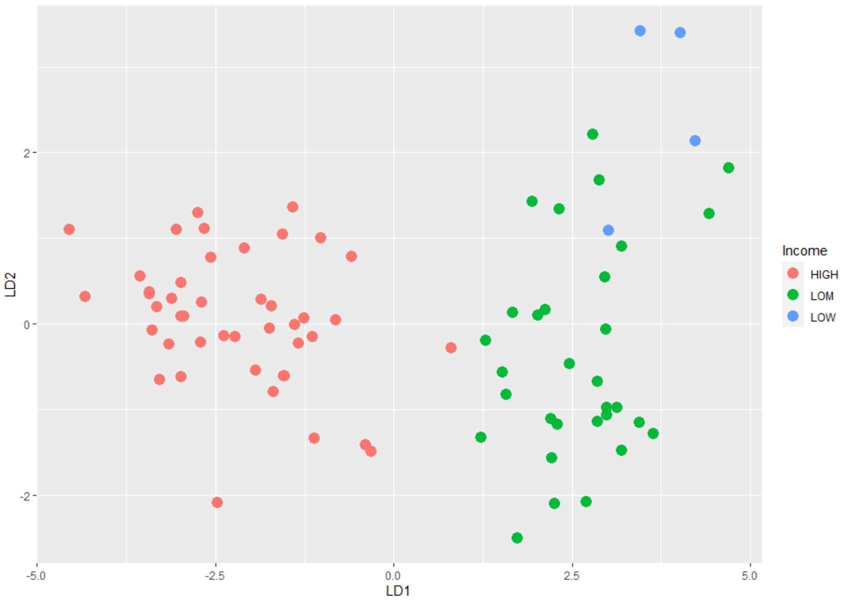


Figure 7: Separation of 3 Groups with Discriminant Analysis.

3 CONCLUSIONS AND LIMITATIONS

3.1 LIMITATIONS

All analysis have their limitations. This analysis is no different. A major limitation faced during the analysis was that data obtained for each variable is not uniform (from different years). Thus, this analysis may not give a true representation of the current state of the world. Also, the data acquired only accounts for 118 countries, which is about 61% of the recognised countries in the world as at 2020 (World Atlas, 2020).

The final limitation faced was the difficulty in interpreting some of the analyses performed. Analyses such as canonical correlation, principal component analysis, and discriminant analysis may be difficult to interpret due to their complexity and may involve a lot of assumptions to be able to be assessed properly (Shiker, 2012).

3.2 CONCLUSION

The analysis in this report was performed to explore and find meaningful information from countries around the world using multivariate analysis methods. After every variable and country was explored through the descriptive statistics in table 1 and the correlation matrix in figure 1, it was revealed that the health variables newbornMortality, childElderPer100, and childPerWoman are highly correlated to one another and the same can be said for the variables, gniPPP, healthSpending, cpi, electricUse, and co2Emissions. This was further confirmed when the canonical correlation analysis between groups of variables was performed.

Also, the PCA showed that HIGH income countries and countries in the ECA geography group tend to have high values for their gniPPP, healthSpending, cpi, electricUse, and co2Emissions variables, while countries in the LOW and LOM income groups and the countries in the SSA geographic region have high values for their newbornMortality, childElderPer100, and childPerWoman variables.

Finally, the differences between the HIGH, LOW and LOM income groups were confirmed with the aid of discriminant analysis. After applying discriminant analysis on the three groups, figure 7 showed that countries in the HIGH income group can be discriminated from the other groups; while countries in the LOM and LOW income groups cannot be discriminated properly. This was also the case in figure 4 when principal component analysis revealed the closeness between the two groups.

4 REFERENCES

Shiker, M. A. (2012). Multivariate statistical analysis. *British Journal of Science*, *6*(1), 55–66.

World Atlas. (2020). *How many countries are there in the world?* https://www.worldatlas.com/articles/how-many-countries-are-in-the-world.html

Transparency International. (2021). *Corruption perceptions index.* https://www.transparency.org/en/cpi/2021?gclid=Cj0KCQjwl7qSBhD-ARIsACvV1X0uiBl8S1-Q-St1ZuQ5oBvRL8vE66S0vnknTk4h-X_pY70StdTzXrsaAmW2EALw_wcB

World Health Organization. (2022). *Diphtheria tetanus toxoid and pertussis (dtp3).* https://www.who.int/data/gho/data/themes/topics/topic-details/GHO/diphtheria-tetanus-toxoid-and-pertussis-(dtp3)

5 APPENDIX

```
#Reading the data into R
ddata=read.csv("countriesdata.csv",row.names=1)

#Extracting the first 14 columns
d1 <- ddata[ ,1:14]

#Plotting the correlation matrix
library(corrplot)
d1_cor <- cor(d1)
corrplot(d1_cor, method = 'number')

#Libraries to aid in the Canonical Correlation analysis
library(CCA)
library(CCP)

#Health and Economic variables Canonical Correlation
model2 <- cc(health, econ)
#Economic and Environmental variables Canonical Correlation
model3 <- cc(econ,env)
#Health and Environmental variables Canonical Correlation
model4 <- cc(health, env)

#Tests for significance of the Canonical Covariates

#Extracting the canoncial covariates
rho <- model2$cor
#Number of observations
n<- dim(health)[1]
#Number of health variables
p <- length(health)
#Number of economic variables
q <- length(econ)
#Function that performs the Wilks Lambda Test
p.asym(rho,n,p,q,tstat = 'Wilks')

#Extracting the canoncial covariates
rho1 <- model4$cor
#Number of observations
n1<- dim(health)[1]
#Number of health variables
p1 <- length(health)
#Number of environmental variables
q1 <- length(env)
#Function that performs the Wilks Lambda Test
p.asym(rho1,n1,p1,q1,tstat = 'Wilks')

#Extracting the canoncial covariates
rho2 <- model3$cor
#Number of observations
n2<- dim(econ)[1]
#Number of economic variables
p2 <- length(econ)
#Number of environmental variables
q2 <- length(env)
```

```
#Function that performs the Wilks Lambda Test
p.asym(rho2,n2,p2,q2,tstat = 'Wilks')

####Modified function from the CCA PACKAGE that helped produce
####the canonical correlation plots
plt.var2 <- function (res, d1, d2, int = 0.5, var.label = FALSE, Xnames = NULL,
                      Ynames = NULL, cex=1)
{
  if (!var.label) {
    plot(0, type = "n", xlim = c(-1.5, 1.5), ylim = c(-1.5, 1.5),
         xlab = paste("Dimension ", d1), ylab = paste("Dimension ",d2))
    points(res$scores$corr.X.xscores[, d1], res$scores$corr.X.xscores[,d2],
    pch = 20, cex = 1.2, col = "red")
    points(res$scores$corr.Y.xscores[, d1], res$scores$corr.Y.xscores[,d2],
    pch = 24, cex = 0.7, col = "blue")
  }
  else {
    if (is.null(Xnames))
      Xnames = res$names$Xnames
    if (is.null(Ynames))
      Ynames = res$names$Ynames
    plot(0, type = "n", xlim = c(-1.5, 1.5), ylim = c(-1.5, 1.5),
         xlab = paste("Dimension ", d1), ylab = paste("Dimension ",
                                                       d2))
    text(res$scores$corr.X.xscores[, d1], res$scores$corr.X.xscores[, d2],
    Xnames, col = "red", font = 2, cex=cex)
    text(res$scores$corr.Y.xscores[, d1], res$scores$corr.Y.xscores[,d2],
    Ynames, col = "blue", font = 3, cex=cex)
  }
  abline(v = 0, h = 0)
  lines(cos(seq(0, 2 * pi, l = 100)), sin(seq(0, 2 * pi, l = 100)))
  lines(int * cos(seq(0, 2 * pi, l = 100)), int * sin(seq(0,2 * pi, l = 100)))
}

#Plots of the three canonical correlation groups with the aid of the CCA package
plt.var2(model2, 1, 2, cex=1.0, var.label = TRUE)
plt.var2(model4, 1, 2, cex=1.0, var.label = TRUE)
plt.var2(model3, 1, 2, cex=0.8, var.label = TRUE)

#PRINCIPAL COMPONENT ANALYSIS#######
#Library that helps to plot the principal components
library(ggbiplot)
#Standardized principal component analysis of the data
pp=princomp(d1, cor=TRUE)
#scree plot of the principal components with the aid of the ggbiplot package
ggscreeplot(pp, type = 'pev')

#Plotting PC1 against PC2
ggbiplot(pp, labels= rownames(d1), var.axes = TRUE)
#Plotting PC1 against PC2 with the geography groupings highlighted
ggbiplot(pp,ellipse=TRUE, groups=dd[,15]) + geom_point(aes(colour=dd[,15]), size = 3)
#Plotting PC1 against PC2 with the income groupings highlighted
ggbiplot(pp,ellipse=TRUE, groups=dd[,16]) + geom_point(aes(colour=dd[,16]), size = 3)

###PATTERN CONFIRMATION (DISCRIMINANT ANALYSIS)######
```

```
#Extracting countries in the HIGH, LOM and LOW income groups
dd5 <- ddata[ddata[,16]=="HIGH"|ddata[,16]=="LOM"|ddata[,16]=="LOW",c(1:16)[-15]]

#Loading the package that will help with the discriminant analysis
library(MASS)
library(ggplot2)
#Discriminant analysis model that separates the three income groups
lda1 <- lda(Income~., data= dd5)
#Summary of the model to view the coefficients of the discriminant functions
summary(lda1)
#Plotting the countries with the two discriminant functions as the axes.
lda.data <- cbind(dd5, predict(lda1)$x)
ggplot(lda.data, aes(LD1, LD2)) + geom_point(aes(color = Income), size = 4)
```