A STUDY ON THE EFFECTS OF HEALTH INITIATIVES AND ECONOMICS ON THE HEALTH STATUS OF A COUNTRY

By:

20-PST-013 I. Albin Praveen Kumar
20-PST-020 Jose Baby

GUIDE:

Dr. Martin Luther William, M.Sc., M.Phil., PhD.

TABLE OF CONTENTS

INTRODUCTION

Health and economy are the most important aspects of a country. These two aspects of a country indicate the quality of life led by its citizens. If a country is poor in these two aspects, it will be evident that citizens are not living happily.

Analyzing the health and economic factors is important for maintaining a good health and economic status of a country and also to improve when health and economic status are poor. This also helps to compare a country's status with other countries to get an idea of where the country stands.

OBJECTIVES OF THE STUDY

01

Identifying latent factors

To identify latent factors among health indicators and to identify latent factors among health initiatives and economic indicators.

02

Ranking the Countries

To rank the countries in terms of health factor and in terms of health initiatives and economic indicators.

03

Finding association between two sets of variables

To find the association between health indicators and health initiatives along with economic indicators.

04

Finding factors influencing important Health indicators

To find which factors in health initiatives influence some important health indicators.

DATA DESCRIPTION

The data was collected from the World Bank open data. This data was taken from the year **2019** as 2020 was an abnormal year. This data contains the Health indicators, Health initiatives and Economic indicators of 192 countries. This data have 16 variables.

Variable	Description	Coded in the Analysis
Crude birth rate	It indicates the number of live births occurring during the year, per 1,000 population estimated at midyear.	CBR
Crude death rate	It indicates the number of deaths occurring during the year, per 1,000 population estimated at midyear.	CDR
Diabetes prevalence	It refers to the percentage of people ages 20-79 who have type 1 or type 2 diabetes. It is calculated by adjusting to a standard population age-structure.	DP
Incidence of tuberculosis	It is the estimated number of new and relapse tuberculosis cases arising in a given year, expressed as the rate per 100,000 population.	TBP
Life Expectancy at birth	It indicates the number of years a newborn infant would live if prevailing patterns of mortality at the time of its birth were to stay the same throughout its life.	LE

Table continues...

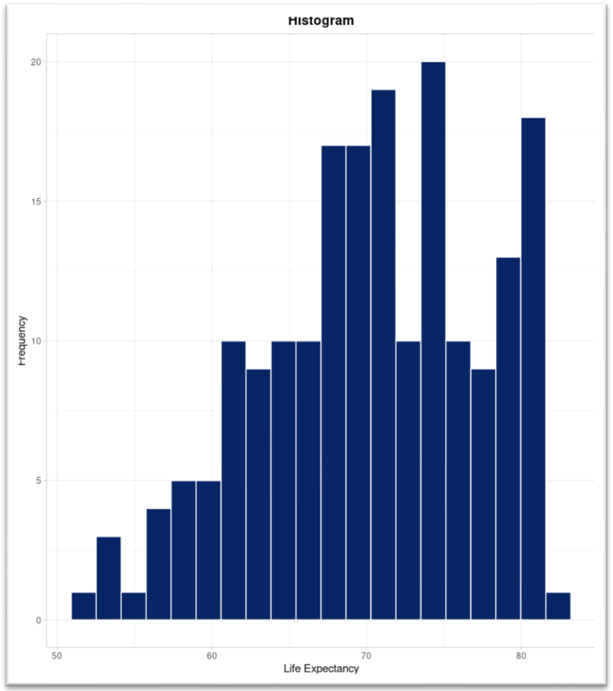
Variable	Description	Coded in the Analysis
Infant mortality rate	It is the number of infants dying before reaching one year of age, per 1,000 live births in a given year.	IMR
Risk of catastrophic expenditure when surgical care	The proportion of population at risk of catastrophic expenditure when surgical care is required. Catastrophic expenditure is defined as direct out of pocket payments for surgical and anaesthesia care exceeding 10% of total income.	RCE
Prevalence of undernourishments	It is the percentage of the population whose habitual food consumption is insufficient to provide the dietary energy levels that are required to maintain a normal active and healthy life.	UND
Level of current health expenditure	It is expressed as a percentage of GDP. Estimates of current health expenditures include healthcare goods and services consumed during each year. This indicator does not include capital health expenditures such as buildings, machinery, IT and stocks of vaccines for emergency or outbreaks.	PGH
General government expenditure on education (current, capital, and transfers)	It is expressed as a percentage of GDP. It includes expenditure funded by transfers from international sources to government. General government usually refers to local, regional and central governments.	PGE

Variable	Description	Coded in the Analysis
Gender parity index	GPI for gross enrollment ratio in primary and secondary education is the ratio of girls to boys enrolled at primary and secondary levels in public and private schools.	GPI
Child immunization, DPT	It measures the percentage of children ages 12-23 months who received DPT vaccinations before 12 months or at any time before the survey. A child is considered adequately immunized against diphtheria, pertussis (or whooping cough), and tetanus (DPT) after receiving three doses of vaccine.	CI_DPT
Inflation	Inflation as measured by the consumer price index reflects the annual percentage change in the cost to the average consumer of acquiring a basket of goods and services that may be fixed or changed at specified intervals, such as yearly. The Lapeyre's formula is generally used.	INF
Unemployment	It refers to the share of the labor force that is without work but available for and seeking employment.	UE
Percentage of people practicing open defecation	This refers to the percentage of the population defecating in the open, such as in fields, forest, bushes, open bodies of water, on beaches, in other open spaces or disposed of with solid waste.	PPOD
Percentage of people using at least basic water services.	This indicator encompasses both people using basic water services as well as those using safely managed water services. Basic drinking water services is defined as drinking water from an improved source, provided collection time is not more than 30 minutes for a round trip. Improved water sources include piped water, boreholes or tube wells, protected dug wells, protected.	PUWS

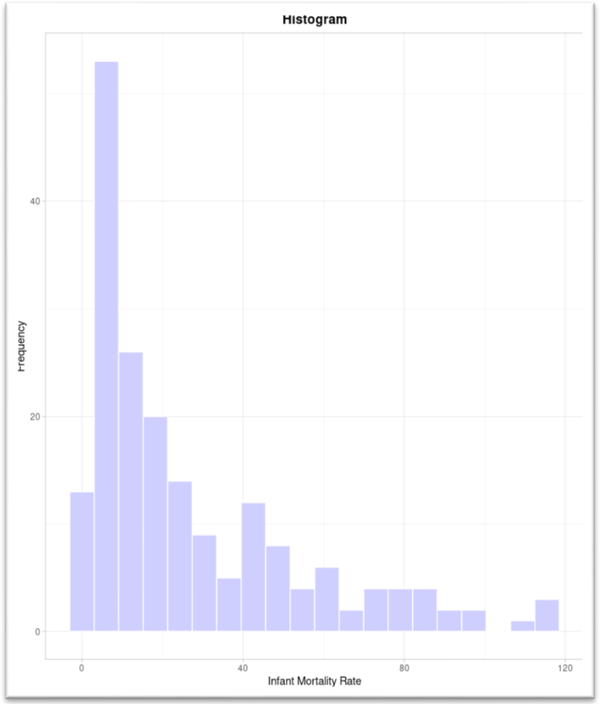
EXPLORATORY DATA ANALYSIS

Univariate Analysis (Health Indicators)

Life Expectancy	
Mean	70.279
Standard Error	0.5162
Median	70.876
Mode	79.142
Standard Deviation	7.153
Sample Variance	51.163
Kurtosis	-0.5361
Skewness	-0.3703
Range	30.699
Minimum	51.201
Maximum	81.9
Sum	13493.535
Count	192



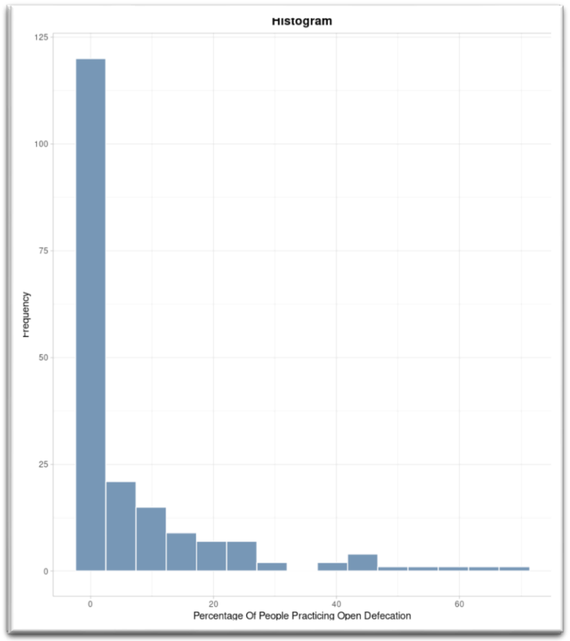
Infant Mortality Rate	
Mean	27.113
Standard Error	1.980
Median	16.25
Mode	3.7
Standard Deviation	27.442
Sample Variance	753.077
Kurtosis	1.2744
Skewness	1.3986
Range	115.5
Minimum	1.7
Maximum	117.2
Sum	5205.6
Count	192



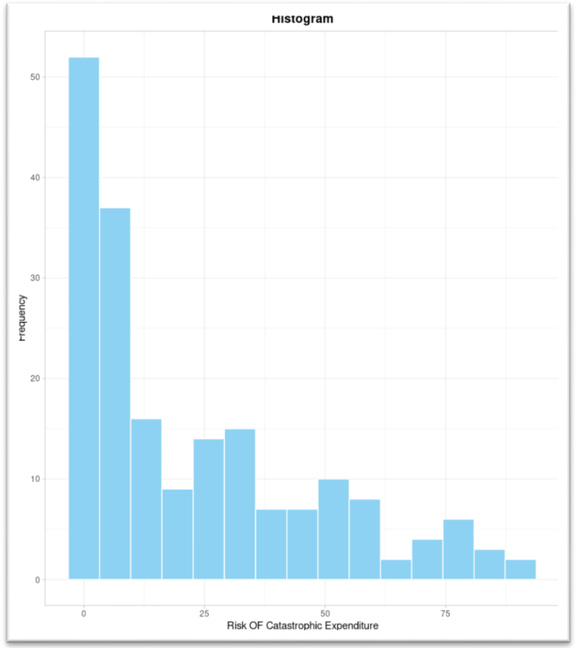
EXPLORATORY DATA ANALYSIS

Health Initiatives and Economic Indicators

Percent Of People practicing OD	
Mean	6.880
Standard Error	0.9355
Median	0.15966
Mode	0
Standard Deviation	12.9621
Sample Variance	168.015
Kurtosis	7.0579
Skewness	2.5891
Range	68.8516
Minimum	0
Maximum	68.8516
Sum	1321.002
Count	192

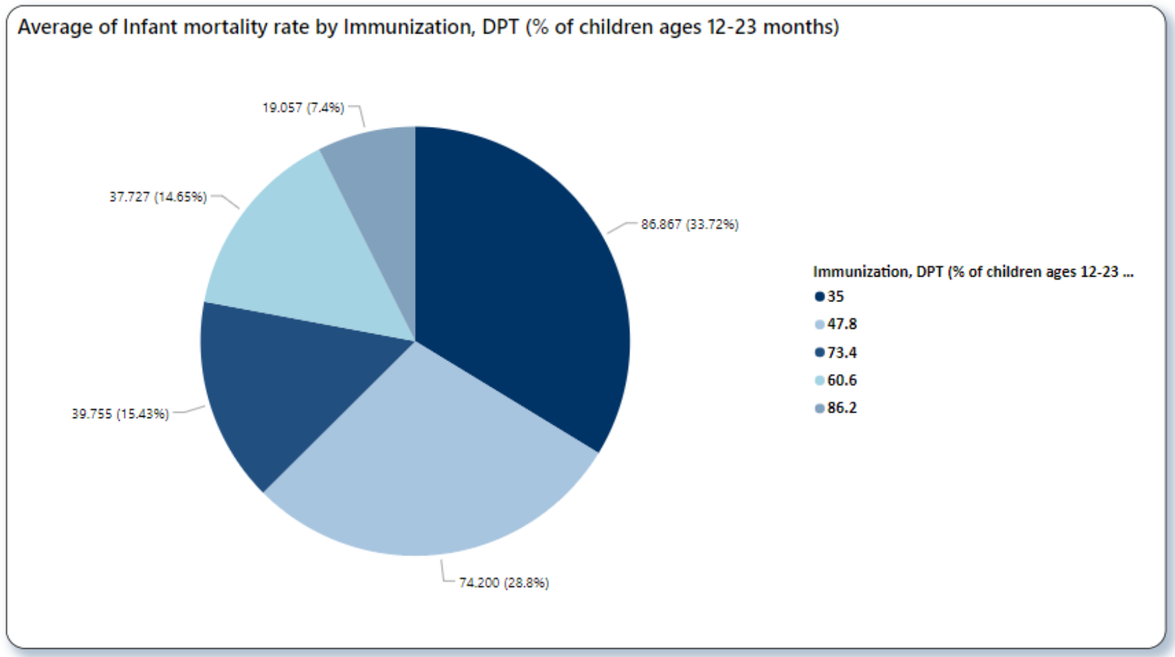


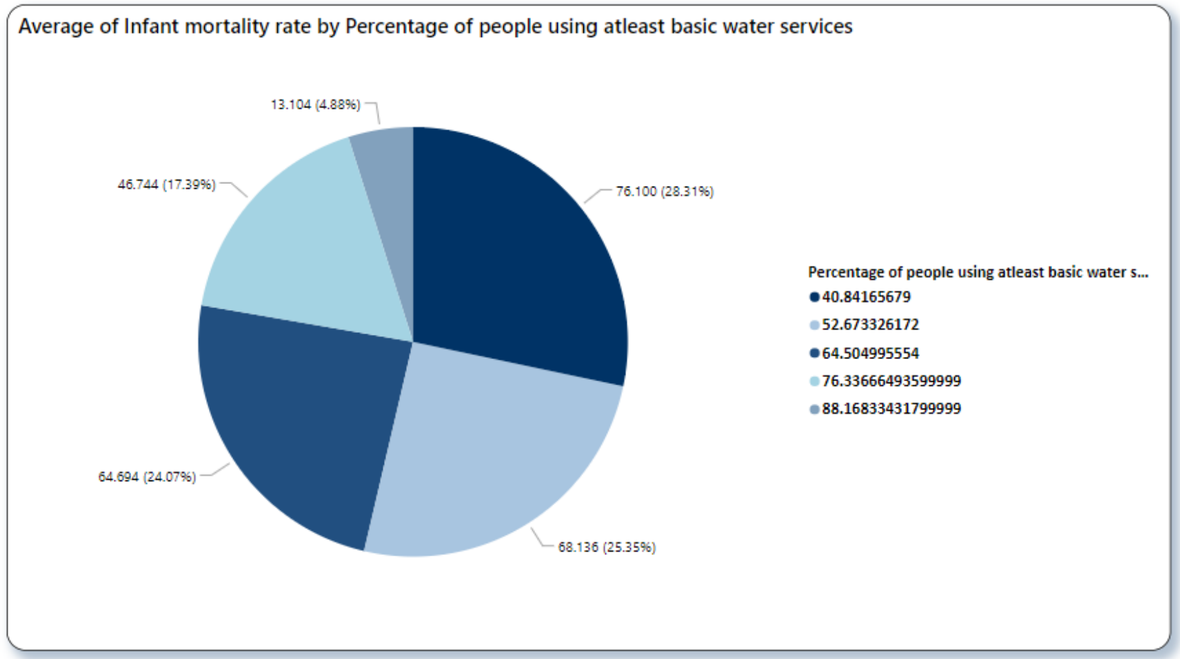
Risk Of Catastrophic Expenditure	
Mean	22.7854
Standard Error	1.7422
Median	11.85
Mode	0.1000
Standard Deviation	24.1406
Sample Variance	582.7693
Kurtosis	0.0606
Skewness	1.0481
Range	90.6
Minimum	0
Maximum	90.6
Sum	4374.8
Count	192



EXPLORATORY DATA ANALYSIS

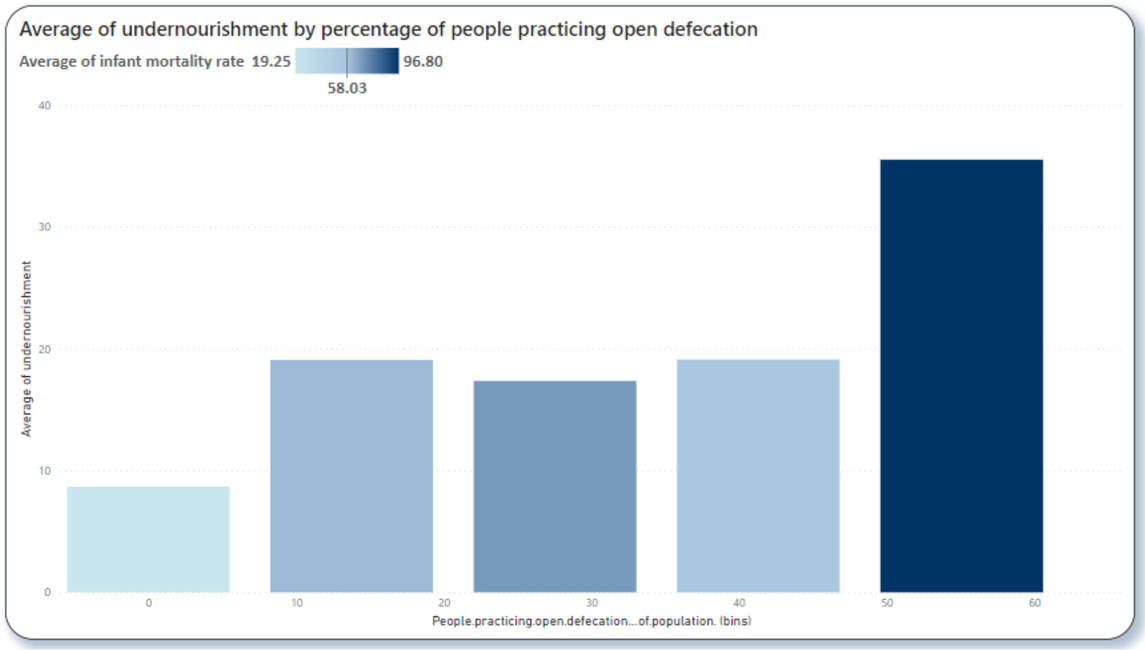
Bivariate Analysis

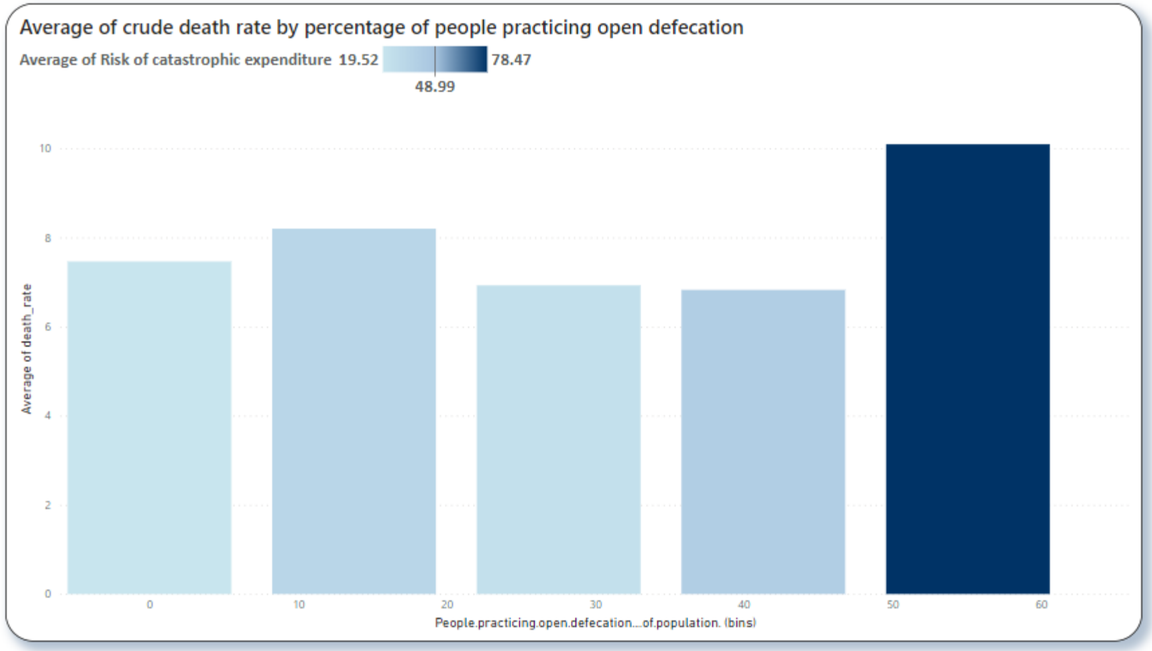




EXPLORATORY DATA ANALYSIS

Bivariate Analysis





Analysis done

The statistical analyses that done for this study

Factor analysis

To identify latent factors and to rank countries.

Canonical Correlation Analysis

To identify association between set of input variables and set of output variables.

Multivariate multiple regression

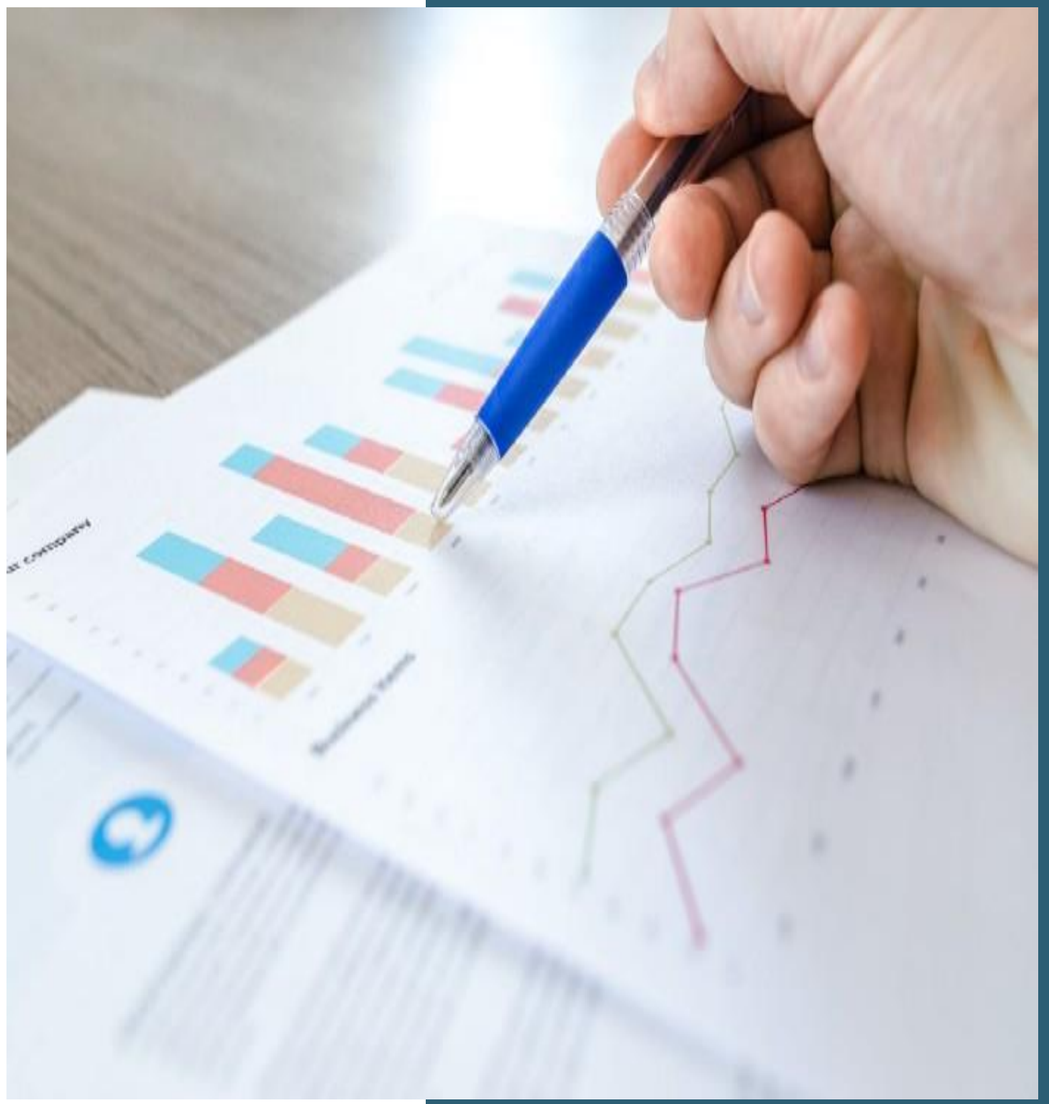
To find which factors in health initiatives influence some important health indicators.

Multivariate outlier detection

Using non parametric approach for multivariate outlier detection to improve model performance

Regression tree

To relate health initiatives with health indicators.



DATA ANALYSIS AND INTERPRETATION

Factor analysis

Health Indicators

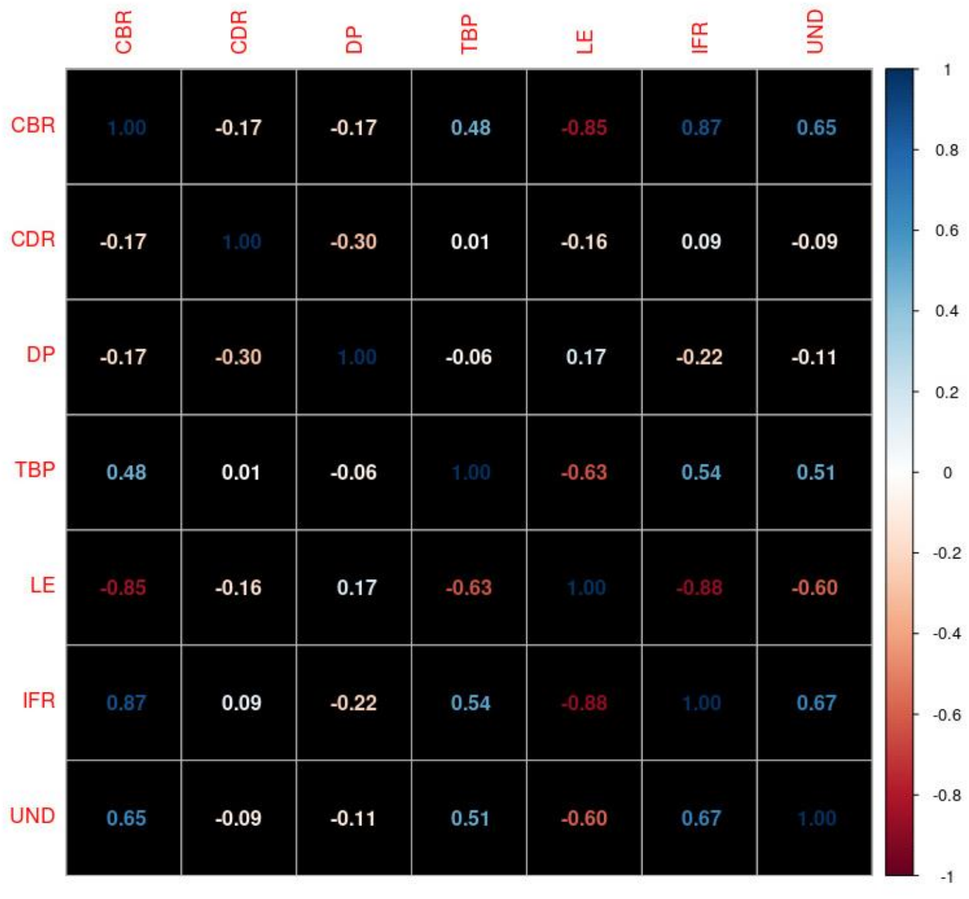
The following are the variables in this group:

- Crude Birth Rate
- Crude Death Rate
- Diabetes Prevalence
- Incidence of Tuberculosis
- Life Expectancy at Birth
- Infant Mortality Rate
- Undernourishment

Factor analysis

Measure of Factorability Of the Data

Checking the Pearson Correlation among the variables.



Kaiser-Mayer-Olkin (KMO) Test

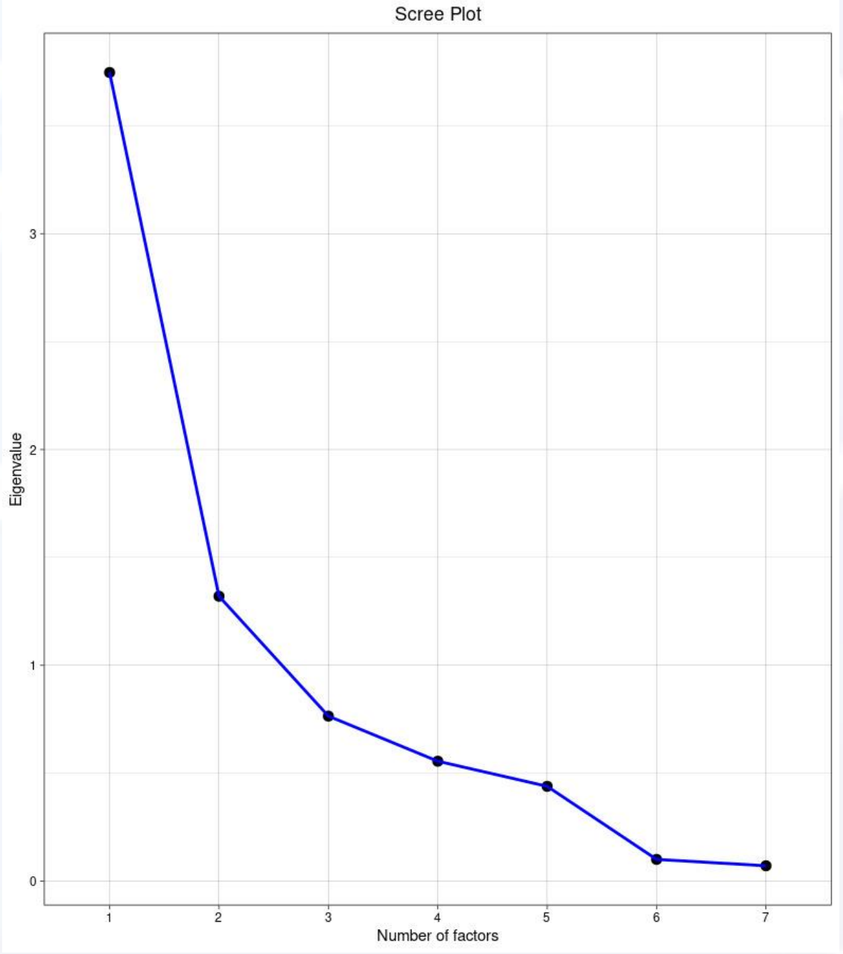
Kaiser_Meyer_Olkin factor adequacy			
CBR	0.68	LE	0.72
CDR	0.16	IMR	0.84
DP	0.5	UND	0.9
TBP	0.76		
Overall MSA		0.71	

Bartlett's Test of Sphericity

Bartlett's Test of Sphericity					
Chi-square	1595.75	p value	$1.44e^{-185}$	D.F	21

Factor analysis

The number of factors to extract



Factor analysis

Conducting the Factor Analysis

Factor Loadings:

Variables	F1	F2	h2	u2
CBR	**0.95**	-0.05	0.9	0.097
CDR	-0.12	**0.99**	1	0.005
DP	-0.16	-0.32	0.13	0.872
TBP	0.58	0.08	0.35	0.652
LE	**-0.9**	-0.27	0.88	0.115
IMR	**0.93**	0.2	0.9	0.101
UND	**0.69**	0	0.48	0.518

Variance Accounted For:

Factors	F1	F2
S.S Loadings	3.43	1.21
Proportion Variance	0.49	0.17
Cumulative Variance	0.49	0.66

Factor scores obtained using the Simple Method

S.no	F1	F2
1	-243.32	-6.285
2	-424.816	-7.981
3	35.782	-8.082
4	59.142	-3.9
5	54.868	-1.521
6	14.209	-7.604

Factor analysis

Health Initiatives and Economic Indicators

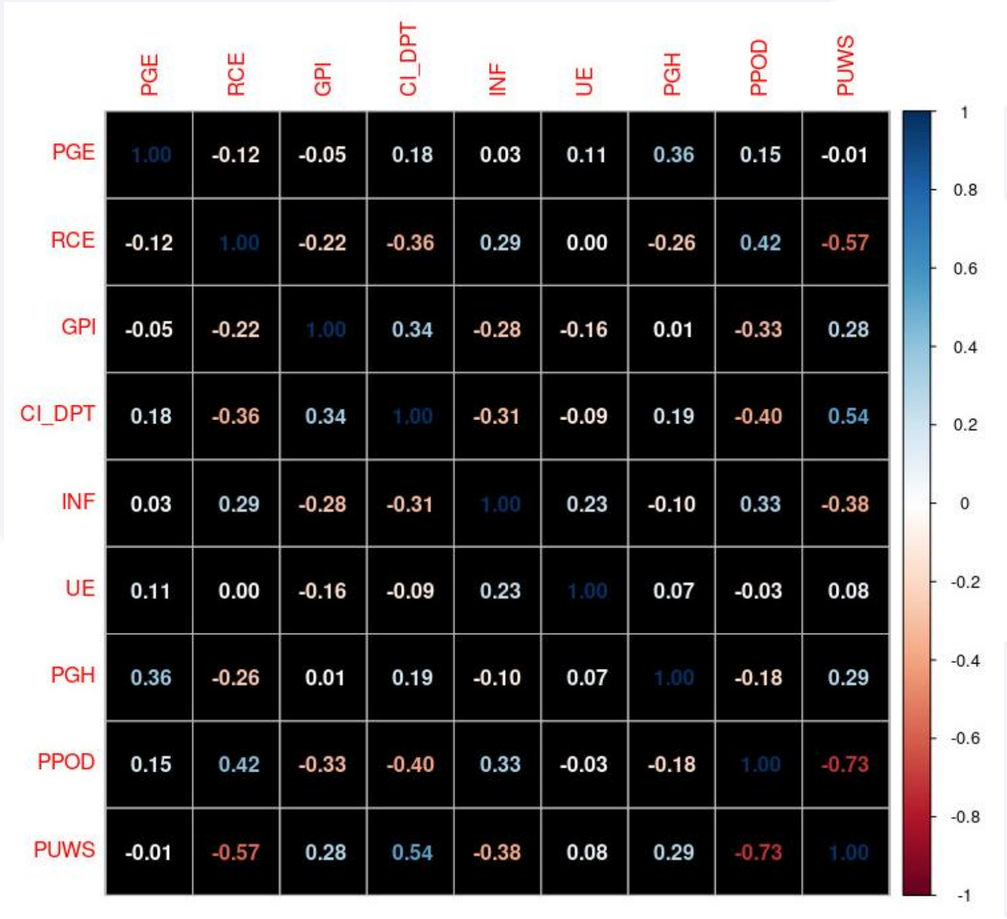
The following are the variables in this group:

1. Crude Birth Rate
2. Crude Death Rate
3. Diabetes Prevalence
4. Incidence of Tuberculosis
5. Life Expectancy at Birth
6. Infant Mortality Rate
7. Undernourishment

Factor analysis

Measure of Factorability Of the Data

Checking the Pearson Correlation among the variables.



Kaiser-Mayer-Olkin (KMO) Test

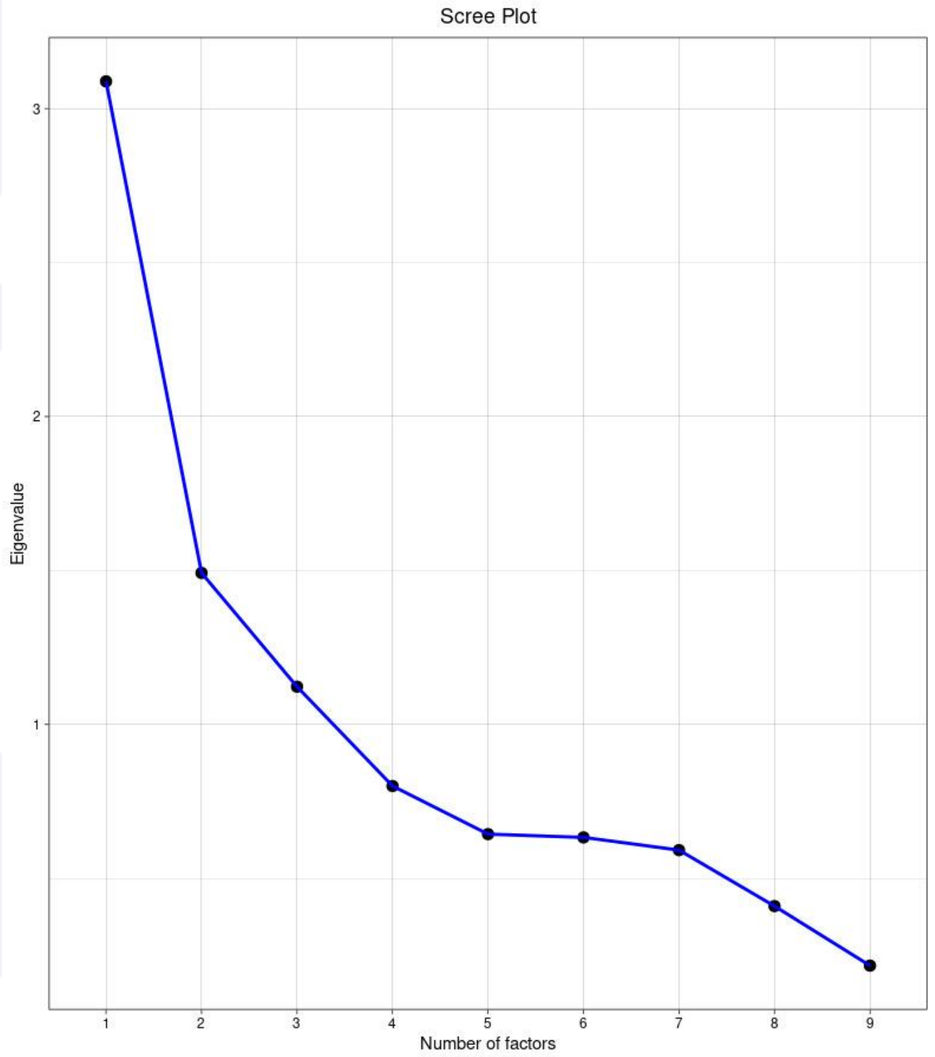
Kaiser_Meyer_Olkin factor adequacy			
PGE	0.44	INF	0.85
RCE	0.85	UE	0.46
GPI	0.81	PPOD	0.74
CI_DPT	0.79	PUWS	0.7
PGH	0.69		
Overall MSA		0.74	

Bartlett's Test of Sphericity

Bartlett's Test of Sphericity					
Chi-square	452.8364	p value	$1.527e^{-73}$	D.F	36

Factor analysis

The number of factors to extract



Factor analysis

Conducting the Factor Analysis

Factor Loadings:

Variables	F3	F2	F1	F4	h2	u2
PGE	-0.07	**0.99**	0.01	0.06	1	0.005
RCE	**-0.58**	-0.16	-0.08	0.05	0.37	0.627
GPI	0.25	-0.04	**0.95**	-0.15	1	0.005
CI_DP	**0.54**	0.23	0.2	-0.17	0.41	0.588
INF	-0.41	-0.01	-0.14	0.36	0.31	0.688
UE	0.04	0.08	-0.07	**0.66**	0.45	0.546
PGH	0.28	0.38	-0.04	0.05	0.23	0.773
PPOD	**-0.75**	0.1	-0.15	-0.03	0.6	0.4
PUWS	**0.96**	0.06	0.06	0.06	0.94	0.06

Variance Accounted For:

Factors	F3	F2	F1	F4
SS loadings	2.44	1.23	1.01	0.63
Proportion Var	0.27	0.14	0.11	0.07
Cumulative Var	0.27	0.41	0.52	0.59

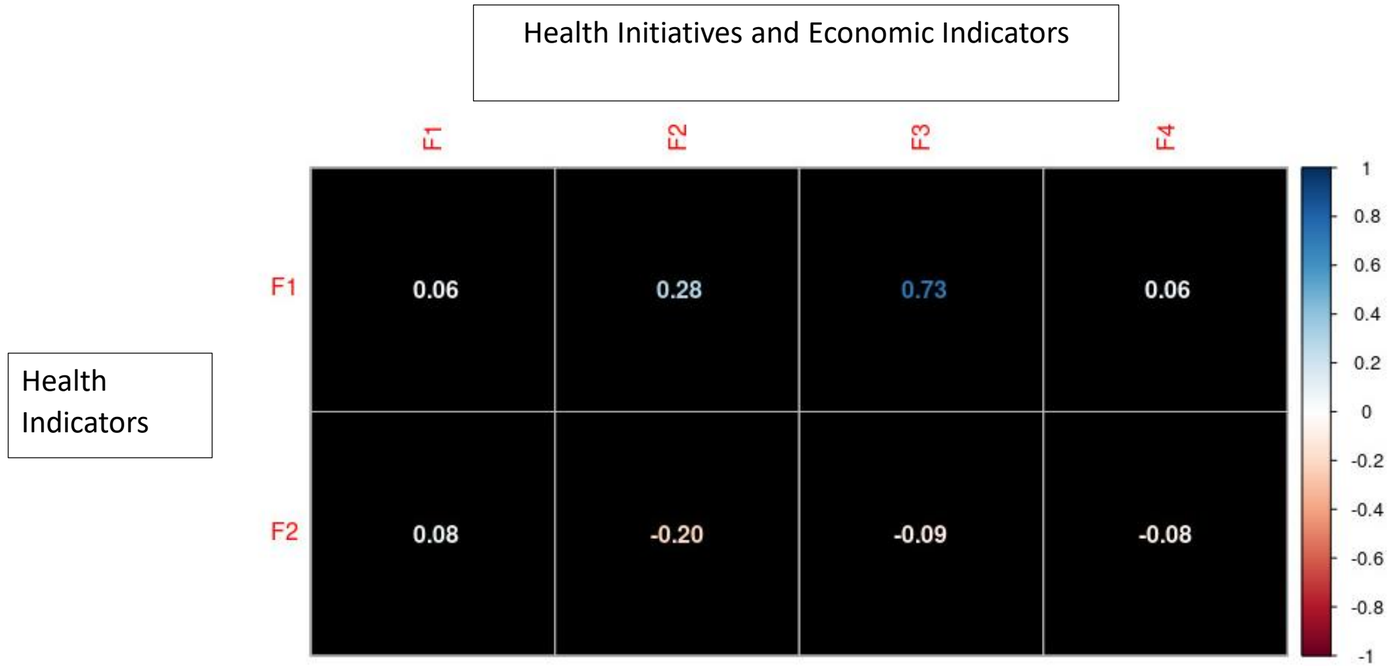
Factor Scores obtained using the Simple Method

S.No	F1	F2	F3	F4
1	0.884	16.456	48.256	10.98
2	1.04	4.3545	32.883	6.93
3	1.019	10.189	180.35	11.47
4	1.006	9.8622	169.13	3.14
5	0.994	7.3777	196.7	2.28
6	1.02	14.312	171.72	9.84

Factor analysis

Spearman Rank Correlation

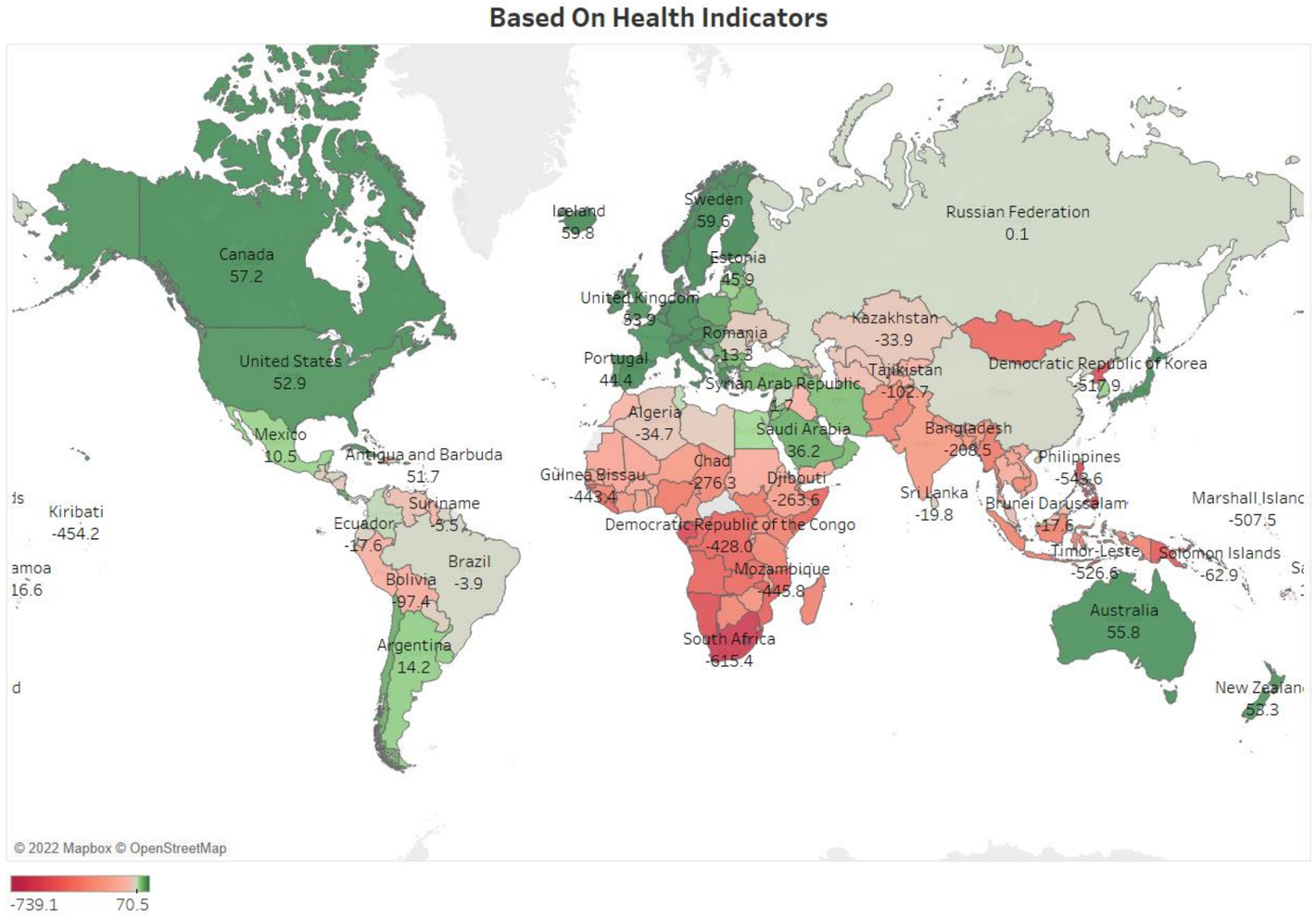
From the Spearman Rank Correlation plot it could be observed that there is high rank correlation between the factor score F1 of the health indicator and factor score F3 of the health initiatives and economic indicator (i.e., 0.73). Thus, we could use these factor scores for ranking countries based on their performances.



Factor analysis

Based on Health Score

The following is a world map, which represents performance of every country based on their health indicators. (We use the Factor score F1 as the latent variable which represents health indicators.)



Countries	Scores
Norway	62.8
Italy	61.4
Finland	61.3
Greece	61.1
Switzerland	60

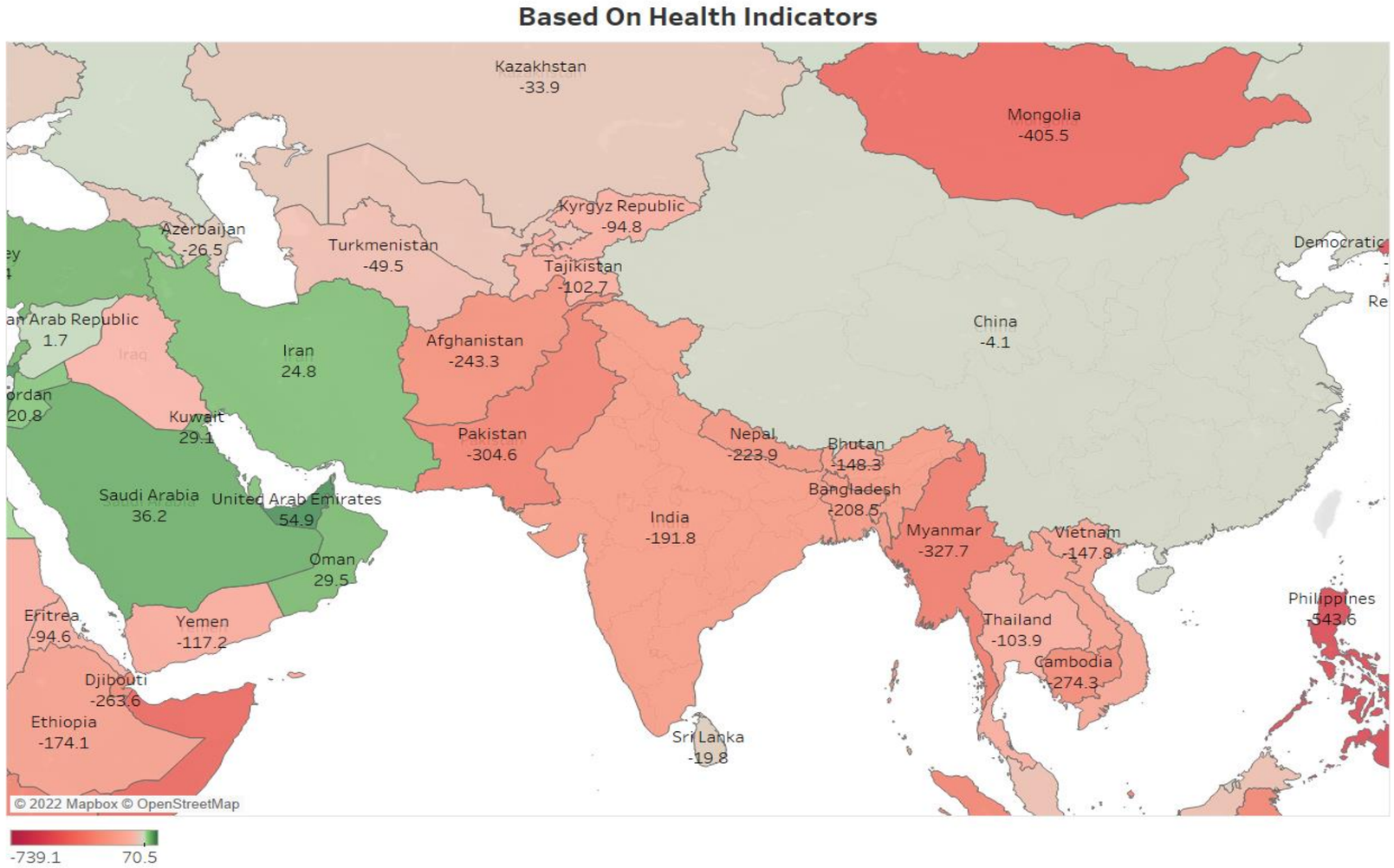
Countries	Scores
Lesotho	-739.1
South Africa	-615.4
Gabon	-545.8
Philippines	-545.6
North Korea	-517.9

The following are the key observations from the map plot:

1. In the **Asian continent**, the **south** and **south-east** countries tend to perform **poorly** in terms of health indicators. **Japan** and **South Korea** are **exceptions** here. They have a good overall health score. In the **Arabian Peninsula** most of the countries except **Yemen** is performing well.
2. Most of the **European countries** except **Ukraine** and **Romania** are performing extremely well. The **Nordic countries** like **Norway, Finland etc.** are one of top performing countries in terms of the health status.
3. In the **African continent, Egypt** is found to be the best performing country.

Factor analysis

India and Its Neighbours

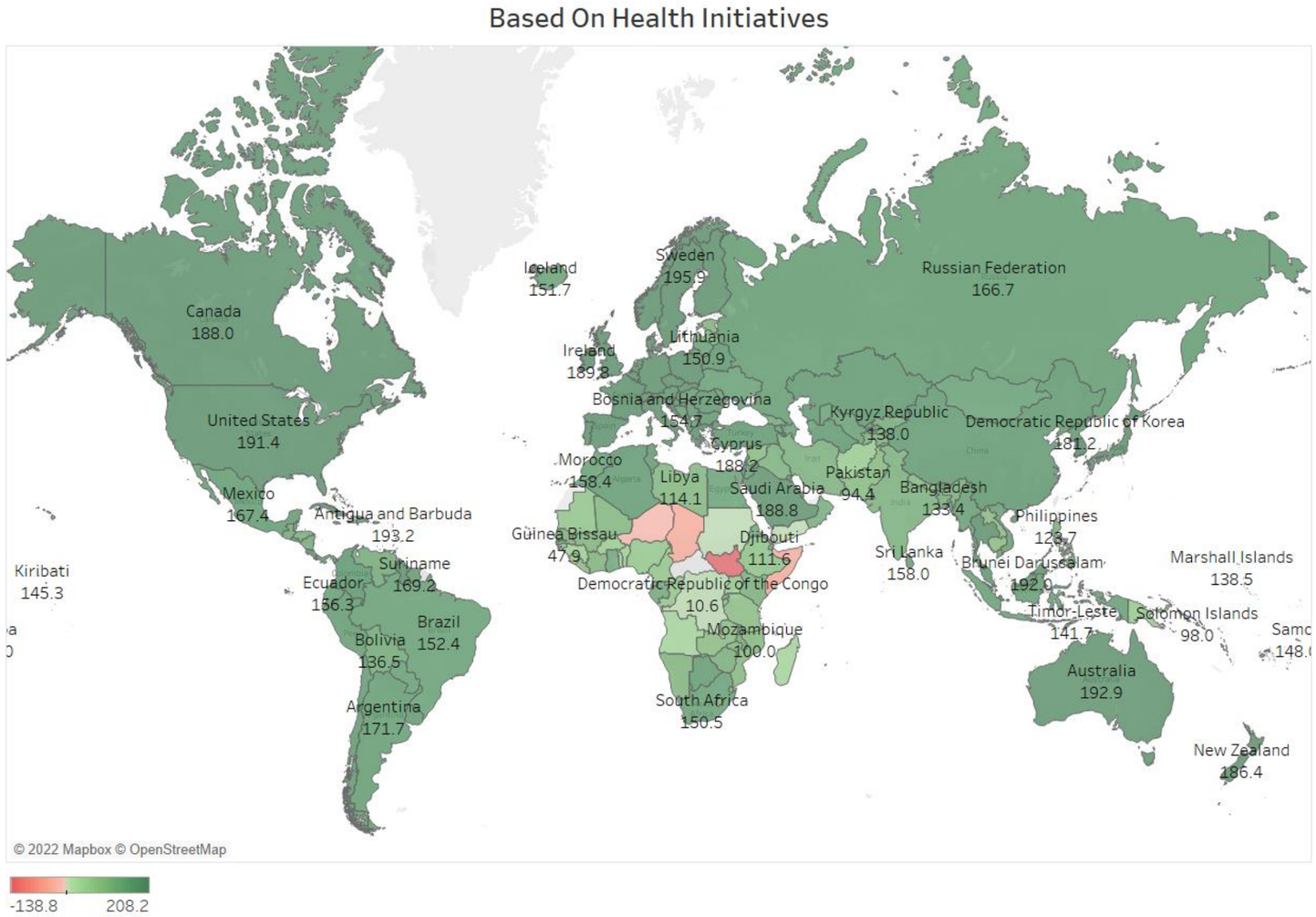


From the Map plot, it could be observed that **India** is comparatively performing better than its neighbouring countries like **Pakistan, Bangladesh, Nepal. (Except Sri Lanka and Bhutan.)** India ranks at the **146th** position among **192** countries in our scoring.

Factor analysis

Based on Health Indicators and Economic indicator Scores

The following is a world map, which represents performance of every country based on the initiatives taken by the respective governments of the country to improve health and economic indicators. (We use the Factor score F3 as the latent variable which represents health initiatives and economic indicators.)



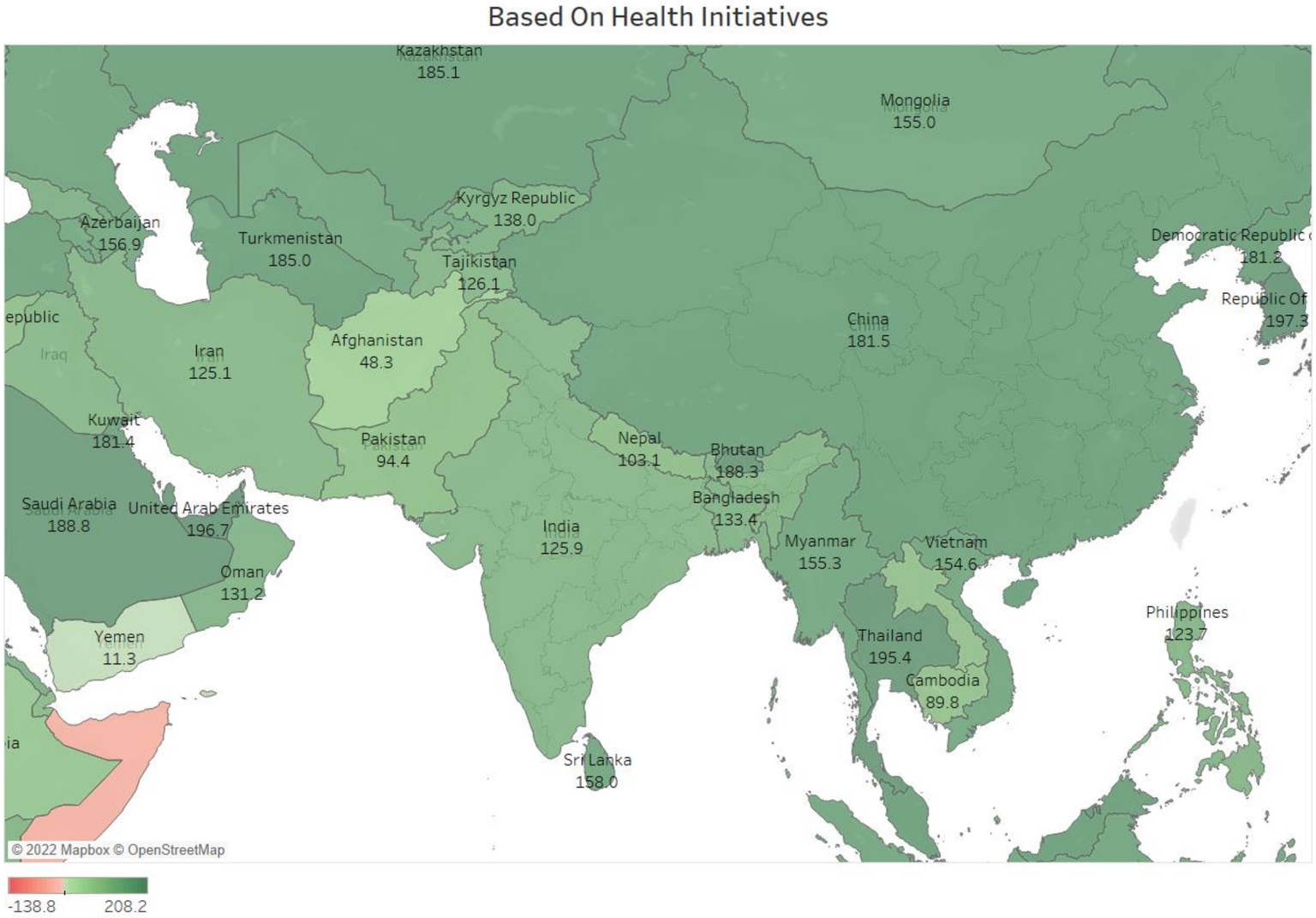
Countries	Scores	Countries	Scores
South Korea	197.3	South Sudan	-138.8
U.A.E	196.7	Chad	-40.9
Belgium	196.6	Somalia	-40.3
Japan	196.2	Niger	-19.8
Denmark	196.1	Sudan	7.5

Observation From the Map plot:

Most of the countries of the western bloc are performing better as compared to countries in the Asian, South American and the African continent. In the Asian continent Japan and South Korea are doing extremely well.

Factor analysis

India and Its Neighbours



In this ranking too, it could be observed that **India** is comparatively performing better than its neighbouring countries like **Pakistan, Bangladesh, Nepal**. (**Except Sri Lanka and Bhutan.**) India ranks at the **138th** position among **192** countries in our scoring.

Canonical Correlation Analysis

Finding association between set of health indicators (Output set) and set of health initiatives along with economic indicators (Input set)

Variables in Output set

- Crude death rate (CDR)
- Diabetes prevalence (DP)
- Incidence of tuberculosis (TBP)
- Life expectancy at birth (LE)
- Infant mortality rate (IMR)
- Prevalence of undernourishments (UND)
- Risk of catastrophic expenditure when surgical care (RCE)

Variables in Input set

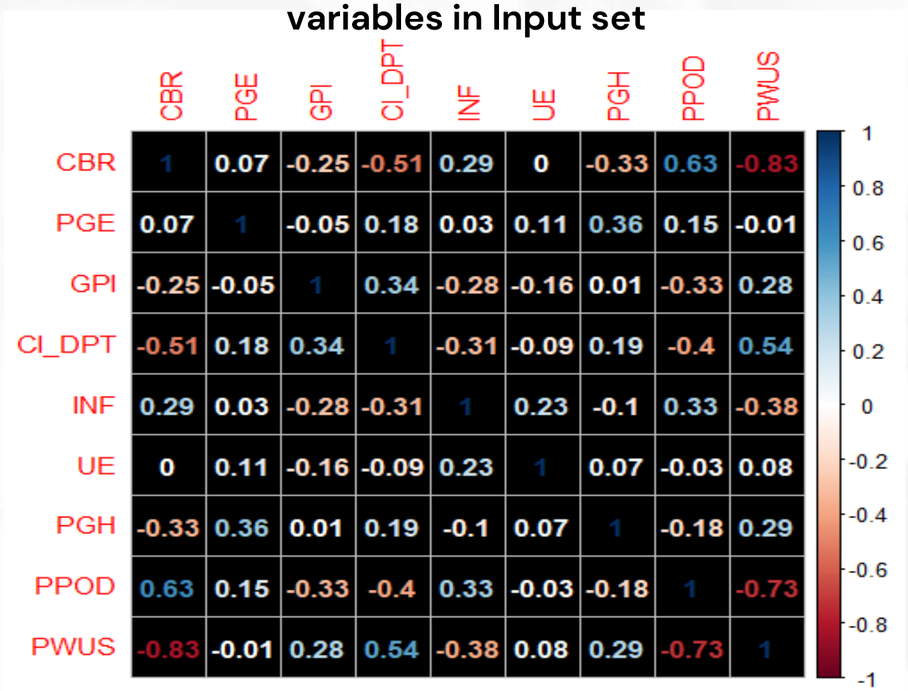
- Crude birth rate (CBR)
- General government expenditure on education (PGE)
- Gender parity index (GPI)
- Child immunization, DPT (CI_DPT)
- Inflation (INF)
- Unemployment(UE)
- Level of current health expenditure (PGH)
- Percentage of people practicing open defecation (PPOD)
- Percentage of people using at least basic water services. (PWUS)

Canonical Correlation Analysis

Pearson correlation

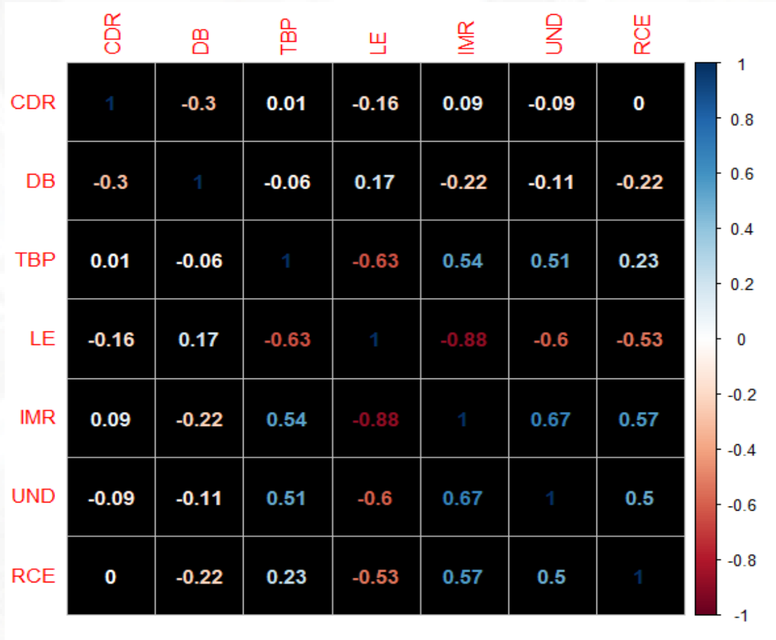
The Pearson correlation among the variables in input set, the correlation among the variables in output set and the correlation between the variables in input set and the variables in output set were found.

Correlation plot for the Pearson correlation among the variables in Input set



- The variables Crude Birth rate and Percentage of people using at least basic water services have high negative correlation.
- The variables Percentage of people practicing open defecation and Percentage of people using at least basic water services have high negative correlation.

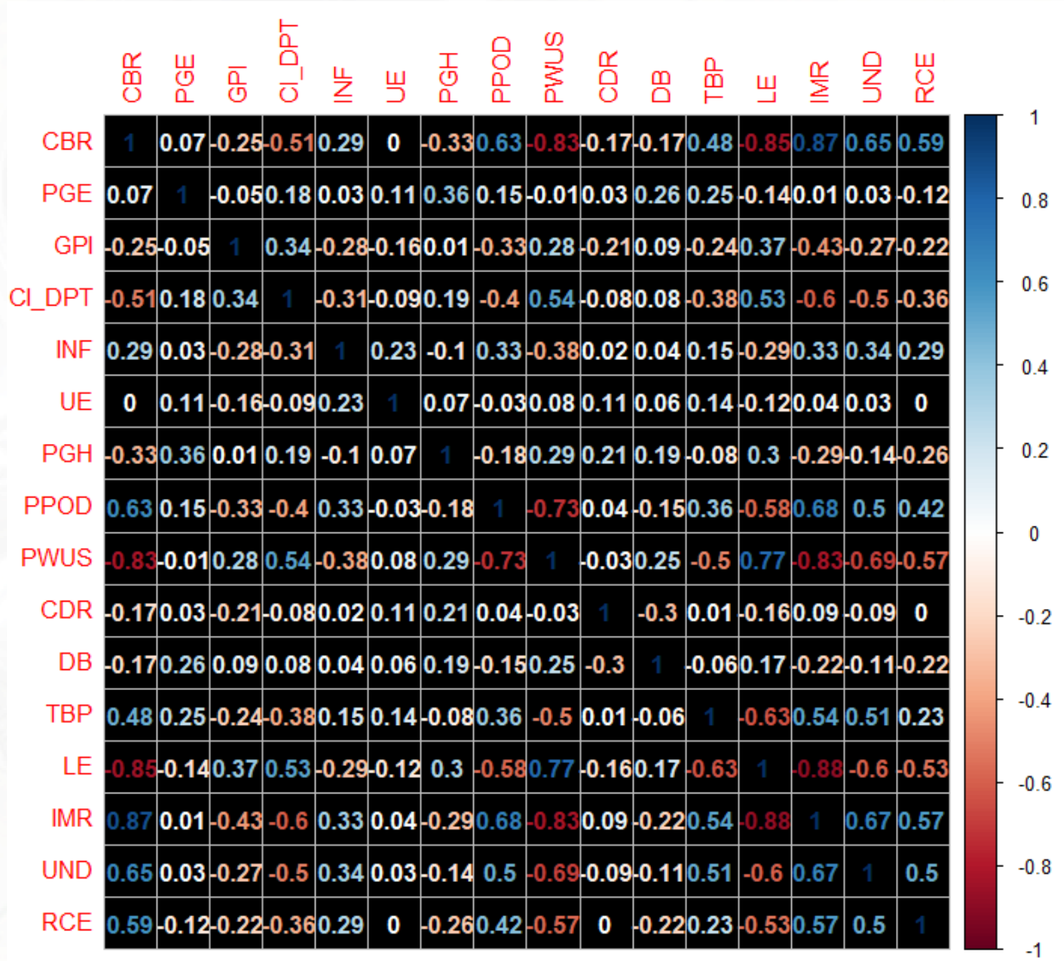
Correlation plot for the Pearson correlation among the variables in Output set



- The variables Infant mortality rate and Life expectancy have high negative correlation.
- The variables Undernourishment and Infant mortality rate have moderate positive correlation.

Canonical Correlation Analysis

Correlation plot for Pearson correlation between the variables in Input set and the variables in Output set



- The variables Crude Birth rate and Life expectancy have high negative correlation.
- The variables Percentage of people using at least basic water services and Infant mortality rate have high negative correlation.

Canonical correlations

The correlations between input and output canonical variates were found

- Seven response variables' canonical variates and seven response variables' canonical variates were created.
- The number of canonical dimensions is generally equal to the number of variables in the smaller set.
- The seven input canonical variates are denotes as, **U1, U2, U3, U4, U5, U6 AND U7.**
- The seven output canonical variates are denoted as **V1, V2, V3, V4, V5, V6 AND V7.**

Variates	Correlations
First input Canonical variate and First output Canonical variate (U1 and V1)	0.95430
Second input Canonical variate and Second output Canonical variate (U2 and V2)	0.61209
Third input Canonical variate and Third output Canonical variate (U3 and V3)	0.52428
Fourth input Canonical variate and Fourth output Canonical variate (U4 and V4)	0.31021
Fifth input Canonical variate and Fifth output Canonical variate (U5 and V5)	0.25387
Sixth input Canonical variate and Sixth output Canonical variate (U6 and V6)	0.12122
Seventh input Canonical variate and Seventh output Canonical variate (U7 and V7)	0.03047

Raw Canonical Coefficients

The raw canonical coefficients can be used to generate the canonical variates.

Coefficients of Input Canonical variates							
	U1	U2	U3	U4	U5	U6	U7
CBR	-0.079	-0.114	-0.005	0.085	0.006	0.006	-0.085
PGE	0.015	0.041	0.580	0.080	0.030	-0.071	0.126
GPI	1.845	-3.760	1.529	1.076	-10.090	-9.575	7.574
CI_DPT	0.005	-0.033	-0.023	-0.033	-0.002	0.039	0.002
INF	0.000	-0.002	-0.015	0.057	-0.073	0.057	0.019
UE	-0.006	0.023	0.061	-0.085	0.026	0.071	0.047
PGH	0.023	0.165	-0.096	0.104	-0.080	-0.075	-0.191
PPOD	-0.003	0.010	-0.044	0.046	0.051	-0.021	0.074
PUWS	0.009	-0.063	-0.016	0.113	0.045	0.008	-0.003

Coefficients of Output Canonical variates							
	V1	V2	V3	V4	V5	V6	V7
CDR	0.091	0.357	0.002	-0.025	0.048	0.183	-0.018
DB	0.017	0.066	0.084	0.195	-0.032	0.065	0.032
TBP	0.001	0.003	0.003	-0.002	0.002	-0.004	0.007
LE	0.063	0.152	-0.189	0.102	0.036	-0.133	0.146
IMR	-0.019	0.023	-0.046	0.044	0.043	-0.023	0.012
UND	-0.006	0.046	-0.002	-0.014	-0.093	-0.027	-0.060
RCE	-0.003	-0.002	-0.011	-0.003	-0.018	0.025	0.042

The raw canonical coefficients are interpreted in a manner analogous to interpreting regression coefficients i.e., for the variable General government expenditure on education (current, capital, and transfers) , a one unit increase in reading leads to a **0.015** increase in the first canonical variate of input set when all of the other variables are held constant.

Canonical loadings

The canonical loadings are correlations between variables and the canonical variates and are important to know about the relevant associations. These canonical variates are a kind of latent variables.

- The variables Crude birth rate, Child immunization DPT, Percentage of people practicing open defecation and Percentage of people using at least basic water services are loaded heavily on the first input canonical variate.
- The variable General government expenditure on education (current, capital, and transfers) is loaded heavily on the third input canonical variate.
- The variable Inflation is loaded heavily on the sixth input canonical variate.

Loadings of input variables on input canonical variates

	U1	U2	U3	U4	U5	U6	U7
CBR	**-0.979**	-0.119	0.053	0.047	-0.058	-0.078	-0.045
PGE	-0.007	0.146	**0.802**	0.284	0.087	-0.066	0.118
GPI	0.379	-0.438	0.060	-0.090	-0.490	-0.538	0.247
CI_DPT	**0.597**	-0.417	0.004	-0.185	0.022	0.154	0.048
INF	-0.344	0.249	-0.060	0.353	-0.526	**0.576**	0.269
UE	-0.031	0.157	0.377	-0.118	0.080	0.525	0.154
PGH	0.391	0.517	0.114	0.306	-0.088	-0.151	-0.399
PPOD	**-0.683**	0.297	-0.180	0.176	0.213	-0.120	0.512
PUWS	**0.877**	-0.284	0.064	0.229	0.186	0.107	-0.145

Canonical loadings

The canonical loadings are correlations between variables and the canonical variates and are important to know about the relevant associations. These canonical variates are a kind of latent variables.

- The variables Life expectancy, Infant mortality rate, Undernourishment and Risk of catastrophic expenditure when surgical care are loaded heavily on the first output canonical variate.
- The variable Crude death rate is loaded heavily on the second output canonical variate.
- The variable Diabetes prevalence is loaded heavily on the fourth input canonical variate.

Loadings of output variables on output canonical variates

	V1	V2	V3	V4	V5	V6	V7
CDR	0.110	**0.694**	-0.002	-0.333	0.330	0.512	-0.155
DB	0.217	-0.005	0.473	**0.821**	-0.213	0.070	0.067
TBP	-0.529	0.318	0.537	-0.209	0.089	-0.380	0.367
LE	**0.909**	-0.111	-0.312	0.126	-0.112	-0.179	0.054
IMR	**-0.955**	0.226	-0.069	0.065	0.170	0.000	-0.003
UND	**-0.710**	0.367	0.008	-0.054	-0.533	-0.242	-0.120
RCE	**-0.644**	0.021	-0.243	-0.117	-0.348	0.406	0.476

Proportion of variation explained by canonical variates

The proportion of variation of input variables explained by Predictor canonical variates were found. This can be useful to determine the variation explained by each of the Predictor canonical variates.

- The first input canonical variate explains **32.6%** of the total variation of input variables.
- The second input canonical variate explains **10.3%** of the total variation of input variables.
- The third input canonical variate explains **9.4%** of the total variation of input variables.
- The fourth input canonical variate explains **4.9%** of the total variation of input variables.
- The fifth input canonical variate explains **6.9%** of the total variation of input variables.
- The sixth input canonical variate explains **10.8%** of the total variation of input variables.
- The seventh input canonical variate explains **6.9%** of the total variation of input variables.
- We can see that **82.2%** of the variation in the input variables is explained by the seven input canonical variates.

Cumulative proportion of variation of input variables explained by Predictor canonical variates	
u1	0.329516
u2	0.432117
u3	0.525962
u4	0.574924
u5	0.644121
u6	0.752973
u7	0.821658

Proportion of variation explained by canonical variates

The proportion of variation of output variables explained by Response canonical variates were found. This can be useful to determine the variation explained by each of the Response canaonical variates.

- The first input canonical variate explains **42.8%** of the total variation of input variables.
- The second input canonical variate explains **11.2%** of the total variation of input variables.
- The third input canonical variate explains **9.6%** of the total variation of input variables.
- The fourth input canonical variate explains **11.9%** of the total variation of input variables.
- The fifth input canonical variate explains **8.7%** of the total variation of input variables.
- The sixth input canonical variate explains **9.5%** of the total variation of input variables.
- The seventh input canonical variate explains **5.8%** of the total variation of input variables.

Cumulative proportion of variation of output variables explained by Response canonical variates	
V1	0.428016
V2	0.539608
V3	0.635757
V4	0.75949
V5	0.846582
V6	0.94181
V7	1

Multivariate multiple regression

Relating the output variables (Dependent variables) with the input variables (Explanatory variables)

From the Canonical Correlation Analysis, it was found that we can relate the four input variables (Crude birth rate, Child immunization DPT, Percentage of people practicing open defecation and Percentage of people using at least basic water services) that are heavily loaded on the first input canonical variate with the four output variables (Life expectancy, Infant mortality rate, Undernourishment and Risk of catastrophic expenditure when surgical care) that are loaded heavily on the first output canonical variate.

Dependent variables

- Life expectancy at birth (LE)
- Infant mortality rate (IMR)
- Prevalence of undernourishments (UND)
- Risk of catastrophic expenditure when surgical care (RCE)

Explanatory variables

- Crude birth rate (CBR)
- Child immunization, DPT (CI_DPT)
- Percentage of people practicing open defecation (PPOD)
- Percentage of people using at least basic water services. (PWUS)

Model coefficients and Significance of independent variables

The model coefficients along with standard error, t statistic value and p value (to find which variables are significantly influencing the dependent variable) were found for each of the dependent variables.

Model 1

Regression equation:

Life expectancy = 66.2778063 - 0.4691743*CBR + 0.0642410*CI_DPT - 0.0008659*PPOD + 0.0863073*PUWS

Interpretation of model coefficients:

- Here the first slope coefficient implies that, if the Crude birth rate increases by 1 unit, the Life expectancy decreases by 0.4691743 units.
- Here the second slope coefficient implies that, if the Child immunization, DPT increases by 1 unit, the Life expectancy increases by 0.0642410 units.
- Here the third slope coefficient implies that, if the Percentage of people practicing open defecation increases by 1 unit, the Life expectancy decreases by 0.0008659 units.
- Here the fourth slope coefficient implies that, if the Percentage of people using at least basic water services increases by 1 unit, the Life expectancy increases by 0.0863073 units.

Dependent variable: Life expectancy				
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	66.2778063	4.3095394	15.379	<2e-16 ***
CBR	-0.4691743	0.0494715	-9.484	<2e-16 ***
CI_DPT	0.0642410	0.0250792	2.562	0.0112 *
PPOD	-0.0008659	0.0301649	-0.029	0.9771
PUWS	0.0863073	0.0364557	2.367	0.0189 *

F Statistic	133 (4 and 187 DF)
p value	< 2.2e-16

Model coefficients and Significance of independent variables

The model coefficients along with standard error, t statistic value and p value (to find which variables are significantly influencing the dependent variable) were found for each of the dependent variables.

Model 2

Regression equation:
Infant mortality rate = 56.04419 + 1.56039*CBR – 0.37890*CI_DPT + 0.29140*PPOD – 0.32090*PUWS

Interpretation of model coefficients:

- If the Crude birth rate increases by 1 unit, the Infant mortality rate increases by 1.56039 units.
- If the Child immunization, DPT increases by 1 unit, the Infant mortality rate decreases by 0.0642410 units.
- If the Percentage of people practicing open defecation increases by 1 unit, the Infant mortality rate increases by 0.29140 units.
- If the Percentage of people using at least basic water services increases by 1 unit, the Infant mortality rate decreases by 0.32090 units.

Dependent variable: Infant mortality rate				
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	56.04419	13.47617	4.159	4.87e-05 ***
CBR	1.56039	0.15470	10.087	< 2e-16 ***
CI_DPT	-0.37890	0.07842	-4.831	2.82e-06 ***
PPOD	0.29140	0.09433	3.089	0.00231 **
PUWS	-0.32090	0.11400	-2.815	0.00540 **

F Statistic	223.8 (4 and 187 DF)
p value	< 2.2e-16

Model coefficients and Significance of independent variables

The model coefficients along with standard error, t statistic value and p value (to find which variables are significantly influencing the dependent variable) were found for each of the dependent variables.

Model 3

Regression equation:

Undernourishment = 49.26957 - 0.25686*CBR - 0.15346*CI_DPT - 0.01989*PPOD - 0.33712*PUWS

Interpretation of model coefficients:

- Here the first slope coefficient implies that, if the Crude birth rate increases by 1 unit, the undernourishment increases by 0.25686 units.
- Here the second slope coefficient implies that, if the Child immunization, DPT increases by 1 unit, the undernourishment decreases by 0.15346 units.
- Here the third slope coefficient implies that, if the Percentage of people practicing open defecation increases by 1 unit, the undernourishment decreases by 0.01989 units.
- Here the fourth slope coefficient implies that, if the Percentage of people using at least basic water services increases by 1 unit, the undernourishment decreases by 0.33712 units.

Dependent variable: Undernourishment				
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	49.26957	9.60810	5.128	7.28e-07 ***
CBR	0.25686	0.11030	2.329	0.02094 *
CI_DPT	-0.15346	0.05591	-2.745	0.00665 **
PPOD	-0.01989	0.06725	-0.296	0.76780
PUWS	-0.33712	0.08128	-4.148	5.09e-05 ***

F Statistic	49.46 (4 and 187 DF)
p value	< 2.2e-16

Model coefficients and Significance of independent variables

The model coefficients along with standard error, t statistic value and p value (to find which variables are significantly influencing the dependent variable) were found for each of the dependent variables.

Model 4

Regression equation:

Risk of catastrophic expenditure when surgical care = 46.70407 + 0.92276*CBR - 0.07775*CI_DPT - 0.01573*PPOD - 0.39760*PUWS

Interpretation of model coefficients:

- Here the first slope coefficient implies that, if the Crude birth rate increases by 1 unit, the Risk of catastrophic expenditure when surgical care increases by 0.92276 units.
- Here the second slope coefficient implies that, if the Child immunization, DPT increases by 1 unit, the Risk of catastrophic expenditure when surgical care decreases by 0.07775 units.
- Here the third slope coefficient implies that, if the Percentage of people practicing open defecation increases by 1 unit, the Risk of catastrophic expenditure when surgical care decreases by 0.01573 units.
- Here the fourth slope coefficient implies that, if the Percentage of people using at least basic water services increases by 1 unit, the Risk of catastrophic expenditure when surgical care decreases by 0.39760 units.

Dependent variable: Risk of catastrophic expenditure when surgical care				
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	46.70407	22.56993	2.069	0.039893 *
CBR	0.92276	0.25909	3.562	0.000468 ***
CI_DPT	-0.07775	0.13135	-0.592	0.554579
PPOD	-0.01573	0.15798	-0.100	0.920787
PUWS	-0.39760	0.19093	-2.082	0.038661 *

F Statistic	27.89 (4 and 187 DF)
p value	< 2.2e-16

Coefficient of determination (R^2) and Adjusted R^2

Model 1

Dependent variable: Life expectancy

- The coefficient of determination (R^2) for this model is 0.7399 which implies that nearly **74%** of the variation in Y1 (Life expectancy) is explained by the regressors (X1, X2, X3 and X4).
- Adjusted R^2 for this model is 0.7343 which implies that nearly **73%** of the variation in Y1 (Life expectancy) is explained by the regressors (X1, X2, X3 and X4).

Coefficient of determination (R^2)	0.7399
Adjusted R^2	0.7343

Model 2

Dependent variable: Infant mortality rate

- The coefficient of determination (R^2) for this model is 0.8272 which implies that nearly **83%** of the variation in Y2 (Infant mortality rate) is explained by the regressors (X1, X2, X3 and X4).
- Adjusted R^2 for this model is 0.8235 which implies that nearly **82%** of the variation in Y2 (Infant mortality rate) is explained by the regressors (X1, X2, X3 and X4).

Coefficient of determination (R^2)	0.8272
Adjusted R^2	0.8235

Coefficient of determination (R^2) and Adjusted R^2

Model 3

Dependent variable: Undernourishment

- The coefficient of determination (R^2) for this model is 0.5141 which implies that nearly **51%** of the variation in Y4 (Undernourishment) is explained by the regressors (X1, X2, X3 and X4).
- Adjusted R^2 for this model is 0.5037 which implies that nearly **50%** of the variation in Y4 (Undernourishment) is explained by the regressors (X1, X2, X3 and X4).

Coefficient of determination (R^2)	0.5141
Adjusted R^2	0.5037

Model 4

Dependent variable: Risk of catastrophic expenditure when surgical care

- The coefficient of determination (R^2) for this model is 0.3736 which implies that nearly **37%** of the variation in Y4 (Risk of catastrophic expenditure when surgical care) is explained by the regressors (X1, X2, X3 and X4).
- Adjusted R^2 for this model is 0.3602 which implies that nearly **36 %** of the variation in Y4 (Risk of catastrophic expenditure when surgical care) is explained by the regressors (X1, X2, X3 and X4).

Coefficient of determination (R^2)	0.3736
Adjusted R^2	0.3602

Test to check whether the predictors jointly contribute to the model

- Determining whether or not to include predictors in a multivariate multiple regression requires the use of multivariate test statistics. The modified hypothesis tests can be used to determine whether a predictor contributes to a model.
- The test statistics calculated is the **Pillai test statistic.** This is a positive valued statistic ranging from 0 to 1. Increasing values means that effects are contributing more to the model; you should reject the null hypothesis for large values.

	Analysis Of Variance Table					
	DF	**Pillai**	**approx F**	**num DF**	**den DF**	**Pr(>F)**
(Intercept)	1	0.99825	26239	4	184	<2.2e-16 ***
CBR	1	0.85141	263.6	4	184	<2.2e-16 ***
CI_DPT	1	0.21202	12.4	4	184	6.184e-09 ***
PPOD	1	0.13428	7.1	4	184	2.134e-0.5 ***
PUWS	1	0.12476	6.6	4	184	5.914e-05 ***
Residuals	187					

Since all the predictors' p values are less than 0.05, it has been found that all the predictors contribute jointly to the four response variables.

Mean Absolute Error (MAE)

The predicted values and the Mean Absolute Errors (MAE) were calculated for each of the models in Multivariate multiple regression. The following are the first six records of the Predicted values of each model in Multivariate multiple regression.

LE	IMR	UND	RCE
61.8387	60.76066	22.66258	41.93789
55.95728	84.30945	31.31561	56.49201
75.36419	6.257409	5.116713	12.05339
77.98418	-2.63368	2.162892	5.706113
76.4693	2.405592	3.001389	8.692718
72.27824	20.68632	7.520004	16.833

The following are the Mean absolute errors of each model in Multivariate multiple regression.

Models	MAE
Model 1	2.86887
Model 2	7.746986
Model 3	5.113802
Model 4	14.25187

Non parametric method of Multivariate Outlier detection

In regression analysis, an outlier is an observation for which the residual is large in magnitude compared to other observations in the data set. The detection of outliers and influential points is an important step of the regression analysis. Now before removing the outliers there is an assumption that needs to be satisfied which is the “errors should follow normality” and if this condition is not satisfied, we cannot use the traditional methods of removing outliers. Thus, a **novel technique** is used, which is a **distribution-free**. The following are the steps involved:

Step 1: Initially a regression model is build and the Mean Square Error (MSE) is found out, which is of the order k x k (k=Number of dependent variables). The square root inverse of MSE is found out. Then we compute the quantity:

$$\boldsymbol{\eta_i = \Sigma^{-\frac{1}{2}} . \varepsilon_i}$$

where i = 1,2, …, n

Each η_i will have the order k x 1

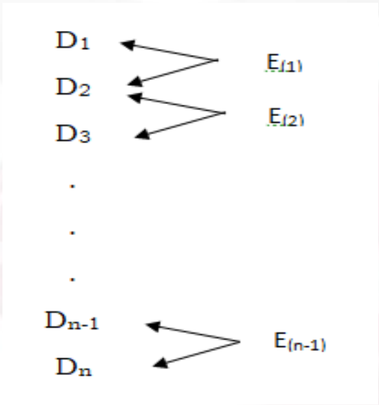
Step 2: Now, a distance function D_i is computed which measures the distance from the origin. It is calculated as

$$\boldsymbol{D_i = |\eta_{i1} - 0| + |\eta_{i2} - 0| + \ldots + |\eta_{ik} - 0|}$$

where i = 1,2, …, n

Non parametric method of Multivariate Outlier detection

Step 3: All D_i are arranged in ascending order.
Step 4: Now we find the successive difference of D_i and denote it as E(j).



Step 5: The 95th Quantile of the E(j)'s is computed and is compared with $E_{(n-1)}$. If 95th Quantile of the E(j) is less than $E_{(n-1),}$ then the observation corresponding to D_n is an outlier.
Step 6: We repeat the Step one to five and rebuild the model after removing the outlier at each iteration. At each iteration, one observation will be removed. The iteration continues till no outlier is detected.

For this data, there were five iterations in which **five outliers were removed.**

Non parametric method of Multivariate Outlier detection

For the outlier free data, Multivariate multiple regression was carried out.

Coefficient of determination (R^2) and Adjusted R^2

Model 1

Dependent variable: Life expectancy

- The coefficient of determination (R^2) for this model is 0.7587 which implies that nearly **76%** of the variation in Y1 (Life expectancy) is explained by the regressors (X1, X2, X3 and X4).
- Adjusted R^2 for this model is 0.7534 which implies that nearly **75%** of the variation in Y1 (Life expectancy) is explained by the regressors (X1, X2, X3 and X4).

Coefficient of determination (R^2)	0.7587
Adjusted R^2	0.7534

Model 2

Dependent variable: Infant mortality rate

- The coefficient of determination (R^2) for this model is 0.8358 which implies that nearly **84%** of the variation in Y2 (Infant mortality rate) is explained by the regressors (X1, X2, X3 and X4).
- Adjusted R^2 for this model is 0.8322 which implies that nearly **83%** of the variation in Y2 (Infant mortality rate) is explained by the regressors (X1, X2, X3 and X4).

Coefficient of determination (R^2)	0.8358
Adjusted R^2	0.8322

Non parametric method of Multivariate Outlier detection

Coefficient of determination (R^2) and Adjusted R^2

Model 3

Dependent variable: Life expectancy

- The coefficient of determination (R^2) for this model is 0.5976 which implies that nearly **60%** of the variation in Y4 (Undernourishment) is explained by the regressors (X1, X2, X3 and X4).
- Adjusted R^2 for this model is 0.5888 which implies that nearly **59%** of the variation in Y4 (Undernourishment) is explained by the regressors (X1, X2, X3 and X4).

Coefficient of determination (R^2)	0.5976
Adjusted R^2	0.5888

Model 4

Dependent variable: Risk of catastrophic expenditure when surgical care

- The coefficient of determination (R^2) for this model is 0.3747 which implies that nearly **37%** of the variation in Y4 (Risk of catastrophic expenditure when surgical care) is explained by the regressors (X1, X2, X3 and X4).
- Adjusted R^2 for this model is 0.3602 which implies that nearly **36%** of the variation in Y4 (Risk of catastrophic expenditure when surgical care) is explained by the regressors (X1, X2, X3 and X4).

Coefficient of determination (R^2)	0.3747
Adjusted R^2	0.361

By comparing the R^2 and adjusted R^2 from the Multivariate multiple regression model for the data with outliers (), the R^2 and adjusted R^2 from the Multivariate multiple regression model for the data without outliers were **slightly improved.**

Test to check whether the predictors jointly contribute to the model

- Determining whether or not to include predictors in a multivariate multiple regression requires the use of multivariate test statistics. The modified hypothesis tests can be used to determine whether a predictor contributes to a model.
- The test statistics calculated is the **Pillai test statistic.** This is a positive valued statistic ranging from 0 to 1. Increasing values means that effects are contributing more to the model; you should reject the null hypothesis for large values.

	Analysis Of Variance Table					
	DF	Pillai	approx F	num DF	den DF	Pr(>F)
(Intercept)	1	0.99832	26749	4	180	<2.2e_16 ***
CBR	1	0.87116	304.3	4	180	<2.2e-16 ***
CI_DPT	1	0.25457	15.4	4	180	7.855e-011 ***
PPOD	1	0.13482	7	4	180	2.869e-0.5 ***
PUWS	1	0.15399	8.2	4	180	4.322e-06 ***
Residuals	183					

Since all the predictors' p values are less than 0.05, it has been found that all the predictors contribute jointly to the four response variables.

Mean Absolute Error (MAE)

The predicted values and the Mean Absolute Errors (MAE) were calculated for each of the models in Multivariate multiple regression. The following are the first six records of the Predicted values of each model in Multivariate multiple regression.

LE	IMR	UND	RCE
76.35159	3.305692	2.680752	9.437591
74.63748	9.942922	4.323096	12.3886
72.99978	15.15525	7.301115	17.97326
74.60409	8.743699	3.439784	12.88566
77.49234	-0.78672	1.78759	7.345219
75.70986	5.920935	3.420322	10.53455

The following are the Mean absolute errors of each model in Multivariate multiple regression.

Models	MAE
Model 1	2.769519
Model 2	7.610648
Model 3	4.494283
Model 4	14.32876

By comparing the Mean Absolute Errors for each dependent variable in the Multivariate multiple regression model for the data with outliers, the Mean Absolute Errors for each dependent variable in the Multivariate multiple regression model for the data without outliers were **slightly reduced.**

.

Regression tree

For each of the the four output variables (Life expectancy, Infant mortality rate, Undernourishment and Risk of catastrophic expenditure when surgical care) that are loaded heavily on the first output canonical variate in Canonical correlation analysis, a regression tree was constructed using the four input variables (Crude birth rate, Child immunization DPT, Percentage of people practicing open defecation and Percentage of people using at least basic water services) that are heavily loaded on the first input canonical variate.

The method used for the construction of the Regression Tree:

1. A large initial regression tree is constructed by using a small value for **cp**, which stands for “complexity parameter.”
2. From the initial construction of the regression tree, various values of cp and the corresponding Errors are obtained.
3. The regression tree is again constructed with above chosen cp value.

Regression tree

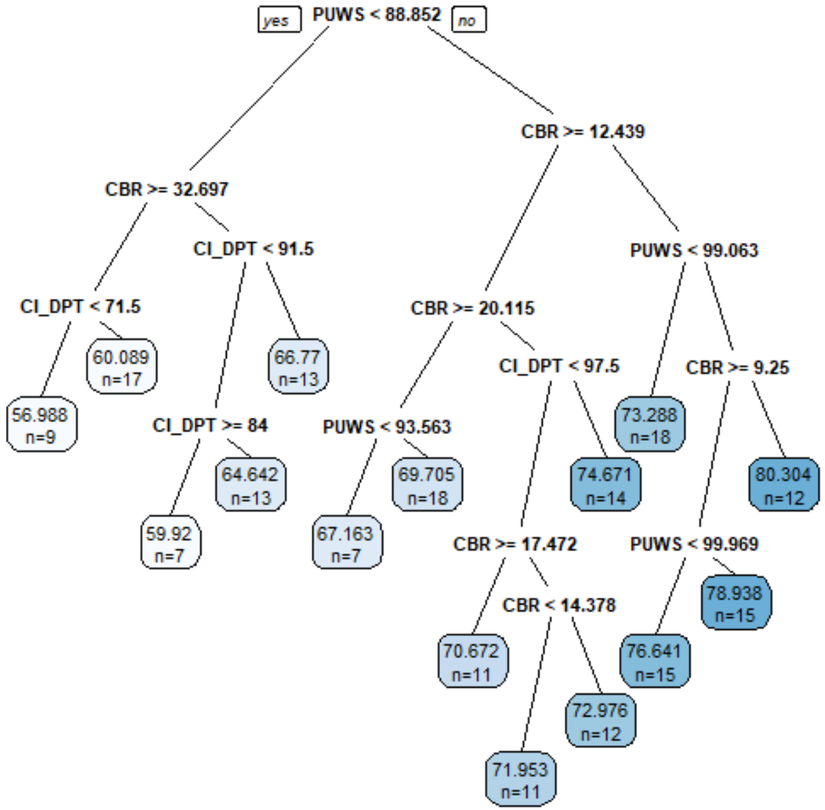
Tree 1 (Dependent variable: Life expectancy at birth)

CP	nsplit	rel error	xerror	xstd
0.5871517	0	1.00000	1.00765	0.087848
0.1021738	1	0.41285	0.43324	0.037602
0.0444206	2	0.31067	0.35001	0.035408
0.0351383	3	0.26625	0.33045	0.035641
0.0232090	4	0.23112	0.30283	0.034699
0.0115267	5	0.20791	0.27956	0.034153
0.0103819	6	0.19638	0.28634	0.036250
0.0077913	7	0.18600	0.28681	0.034847
0.0057907	8	0.17821	0.28455	0.034793
0.0055467	9	0.17242	0.28652	0.034433
0.0040490	10	0.16687	0.28363	0.034209
0.0033334	11	0.16282	0.28206	0.034579
0.0025066	12	0.15949	0.28563	0.034793
0.0006141	13	0.15698	0.28135	0.033707
0.0000000	14	0.15637	0.27960	0.033538

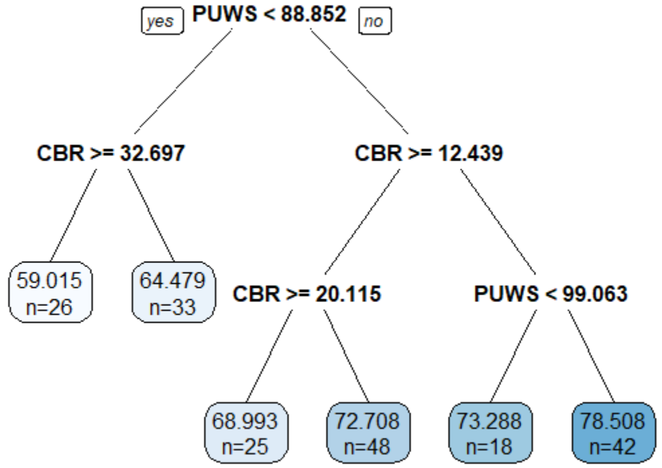
Regression tree

Tree 1 (Dependent variable: Life expectancy at birth)

Before Pruning



After Pruning



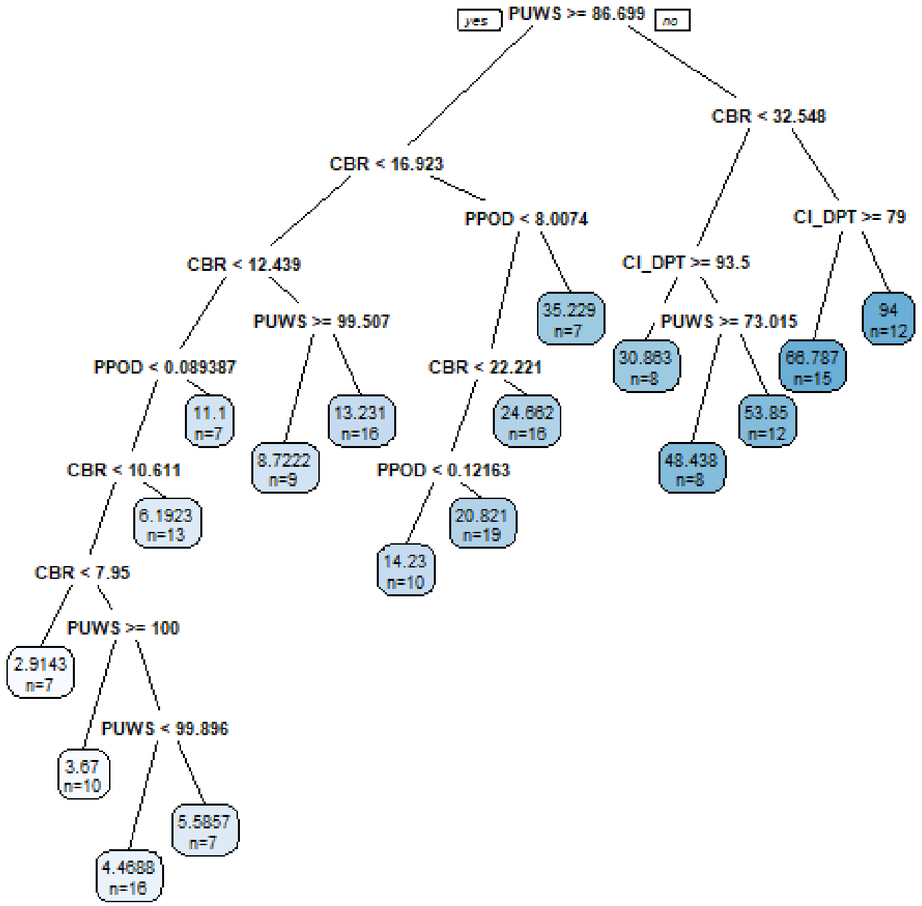
The above tree has **six terminal nodes.** Each terminal node shows the predicted Life expectancy in that node along with the number of observations from the original dataset that belong to that node.

In the original dataset there were 26 countries with PUWS less than 88.852%, CBR greater than or equal to 32.697 and the countries' average LE was 59.015. In this way, all the terminal nodes can be interpreted.

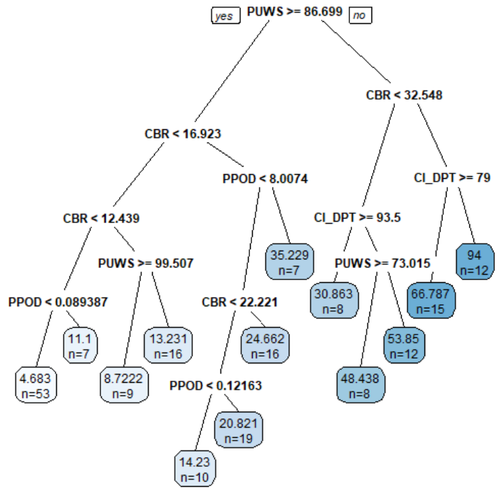
Regression tree

Tree 2 (Dependent variable: Infant mortality rate)

Before Pruning



After Pruning



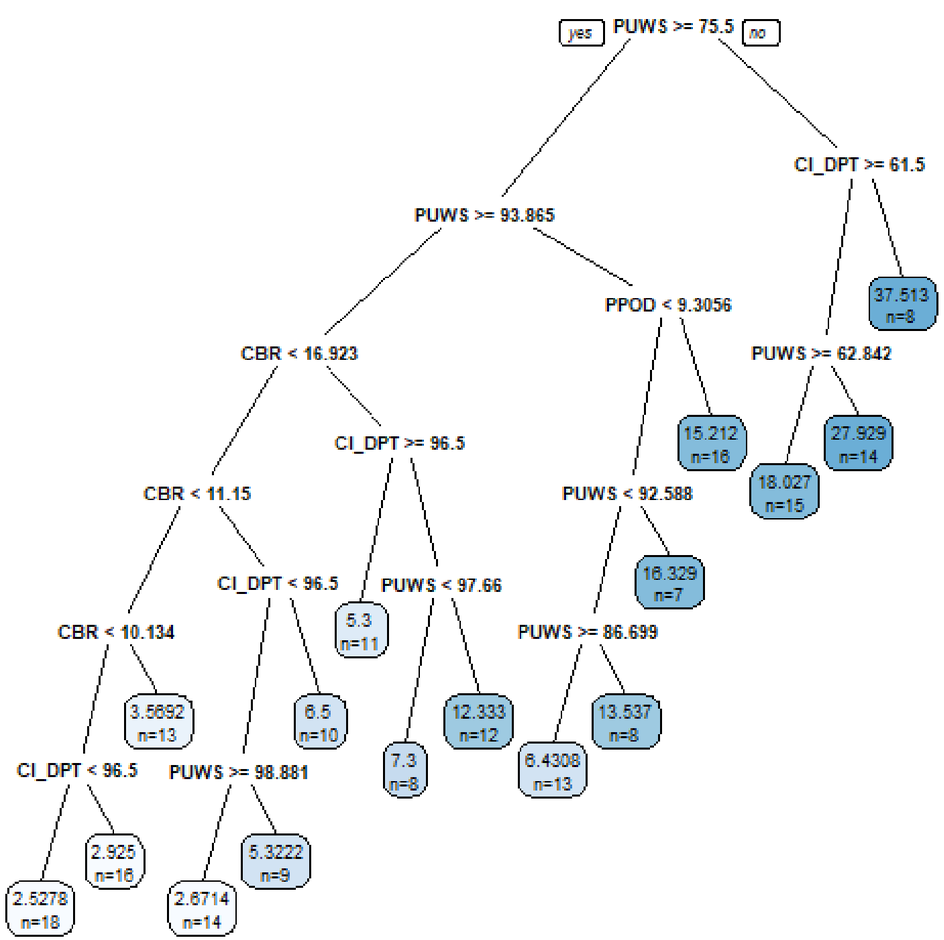
The above tree has **thirteen terminal nodes.** Each terminal node shows the predicted Infant mortality rate in that node along with the number of observations from the original dataset that belong to that node.

In the original dataset there were 9 countries with PUWS greater than or equal to 86.659%, CBR less than 32.697 and CI_DPT greater than or equal to 79% and the countries' average IMR was 94. In this way, all the terminal nodes can be interpreted.

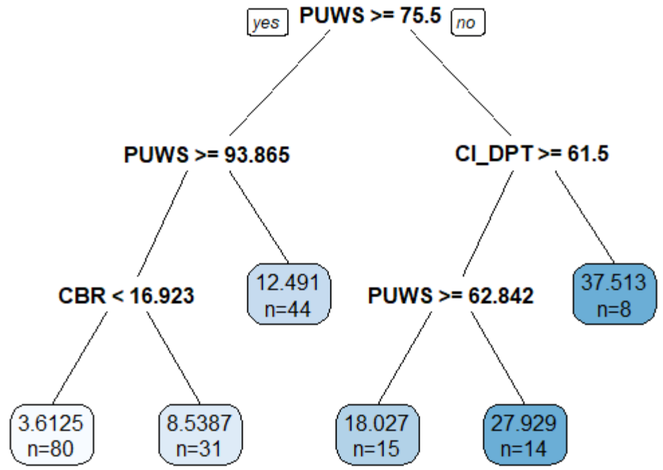
Regression tree

Tree 3 (Dependent variable: Undernourishment)

Before Pruning



After Pruning



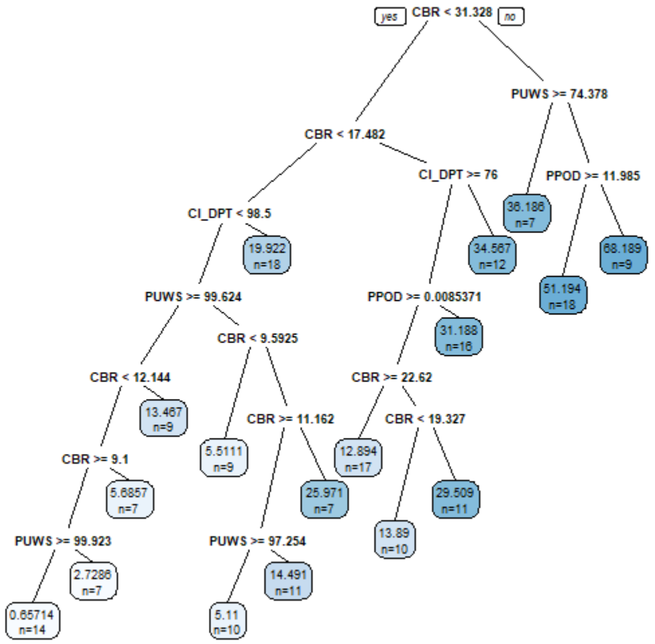
The above tree has **six terminal nodes.** Each terminal node shows the predicted undernourishment in that node along with the number of observations from the original dataset that belong to that node.

In the original dataset there were 8 countries with PUWS greater than or equal to 75.5%, CI_DPT greater than or equal to 61.5 and the countries' average UND was 37.513. In this way, all the terminal nodes can be interpreted.

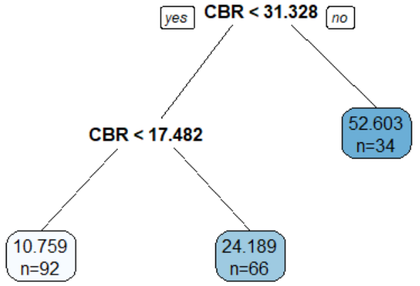
Regression tree

Tree 4 (Dependent variable: Risk of catastrophic expenditure when surgical care)

Before Pruning



After Pruning



- The above tree has **three terminal nodes.** Each terminal node shows the predicted risk of catastrophic expenditure in that node along with the number of observations from the original dataset that belong to that node.
- In the original dataset there were 92 countries with CBR less than 31.328 and CBR less than 17.482 and the countries' average RSE was 10.759. In this way, all the terminal nodes can be interpreted.

Mean Absolute Error (MAE)

The predicted values and the Mean Absolute Errors (MAE) were calculated for each pruned regression tree. The following are the first six records of the Predicted values of each pruned regression tree.

LE	IMR	UND	RCE
64.642	53.85	18.02667	52.60294
56.98778	94	37.5125	52.60294
73.28778	4.683019	3.6125	10.7587
80.30433	4.683019	3.6125	10.7587
76.6412	4.683019	3.6125	10.7587
72.4867	13.23125	3.6125	10.7587

The following are the Mean absolute errors of each pruned regression tree.

Models	MAE
Tree 1	2.209344
Tree 2	6.28956
Tree 3	4.820784
Tree 4	11.73056

CONCLUSION

- The 192 countries were ranked according to the Basic health scores. To accomplish this, Factor analysis was used. It was found that **Norway** is performing best in terms of Basic health score and **Lesotho** is performing worst in terms of Basic health score.
- The 192 countries were ranked according to the Health initiatives. To accomplish this, Factor analysis was used. It was found that **South Korea** is performing best in terms of Health initiatives and economic indicators score and **South Sudan** is performing worst in terms of Health initiatives and economic indicators score.
- The output variables (Health indicators) that are loaded (heavily correlated) with the first canonical output variate suggest a ‘separation’ of the output variables and thereby enables separate ‘treatment / further analysis’ of the subset, the subset containing fewer variables than the original set. The four variables that are heavily loaded in the first canonical output variate were Life expectancy, Infant mortality rate, undernourishment and Risk of catastrophic expenditure. So, these four variables alone can be used as dependent variables in further analyses.
- The input variables that are loaded (heavily correlated) with the first canonical input variate suggest not only a ‘separation’ but also in reducing the number of input variables that are ‘really’ relevant or significant for the further analysis to relate the inputs to the outputs. The four variables that are heavily loaded in the first canonical output variate were Crude birth rate, Infant mortality rate, Percentage of people practicing open defecation and Percentage of people using at least basic water services. So, these four variables alone can be used as independent variables in further analyses.

CONCLUSION

- The independent variables that were found from Canonical correlation analysis were related to the dependent variables that were found from Canonical correlation analysis and the variables that influence each of the dependent variables significantly were observed and also whether all the independent variables jointly contribute to the four dependent variables were found.
- For finding the important variables that influences each of the four dependent variables and to find the predicted values for each dependent variables, **Regression trees** were constructed.
- The optimal CP values were found for which x-error is minimum.
- The trees were pruned further to improve the performance by choosing optimal CP values.
- For detection of multivariate outliers, a non-parametric approach was used.
- For this data, there were five iterations in which **five outliers were removed** and for the outlier free data, Multivariate multiple regression was carried out.
- By comparing the R^2 and adjusted R^2 from the Multivariate multiple regression model for the data with outliers, the R^2 and adjusted R^2 from the Multivariate multiple regression model for the data without outliers were **slightly improved.**
- By comparing the Mean Absolute Errors for each dependent variable in the Multivariate multiple regression model for the data with outliers ,the Mean Absolute Errors for each dependent variable in the Multivariate multiple regression model for the data without outliers were **slightly reduced.**

BIBLIOGRAGHY

BOOKS:

- Anderson, Theodore W., An Introduction to Multivariate Statistical Analysis, Third Edition, John Wiley and Sons.
- Johnson and Wichern, Applied Multivariate Statistical Analysis, Sixth Edition, Pearson India.

WEB SOURCES:

- https://data.library.virginia.edu
- https://www.analyticsvidhya.com
- https://towardsdatascience.com
- https://stats.oarc.ucla.edu

THANK YOU