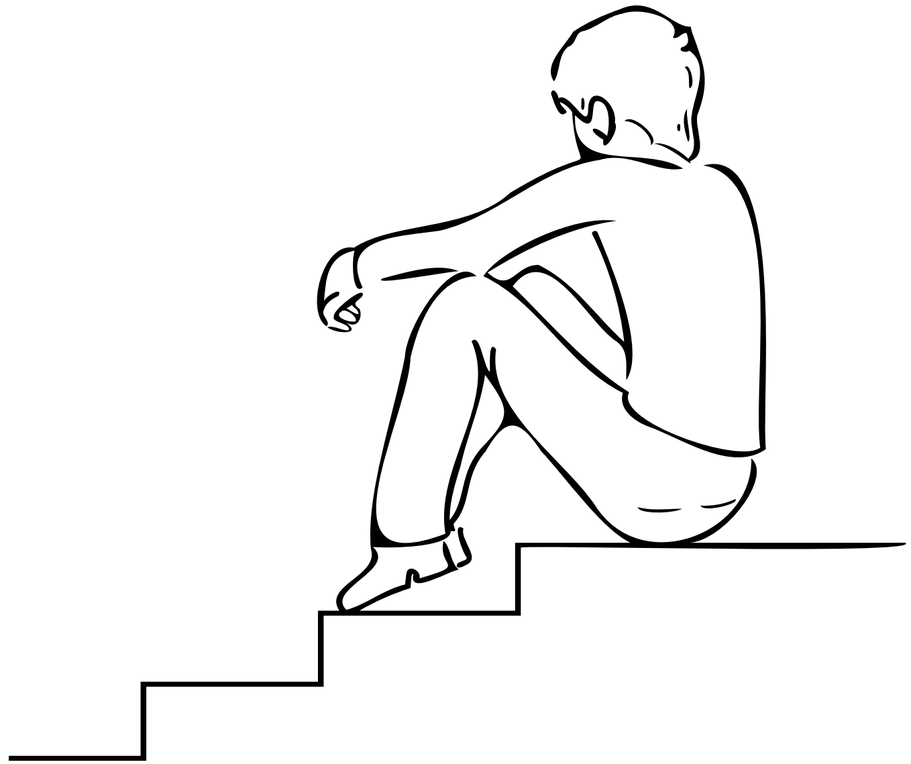
TOMMASO PREMOLI - NR. 34221A

A Worldwide Exploration of Depression

Understanding the Factors

Contents

- Research objective
- Dataset and variables
- Exploratory data analysis (EDA)
- Unsupervised learning
  - Principal component analysis (PCA)
  - Clustering
- Supervised learning
  - Linear ands stepwise regression
  - Ridge and Lasso Regression
  - Decision tree
- Conclusion



Research Objective

Analysing the socio-economic, demographic and health characteristics of countries globally.

Identify common patterns and distinct geographical clusters based on the similarity of individual observations.

Create a predictive model to estimate the level of depression in the population of each country.

Identify the most significant socio-economic, demographic and health variables that have a significant impact on the mental health index.

Dataset

158 countries	Year 2019

Depression

Anxiety disorder

Bipolar disorder

Drug use disorder

Eating disorder

Alcohol disorder

Obesity

Suicide rate

GDP per capita

Health expenditure

Income

Unemployment

Urbanisation

Internet access

Average years of schooling

Phones

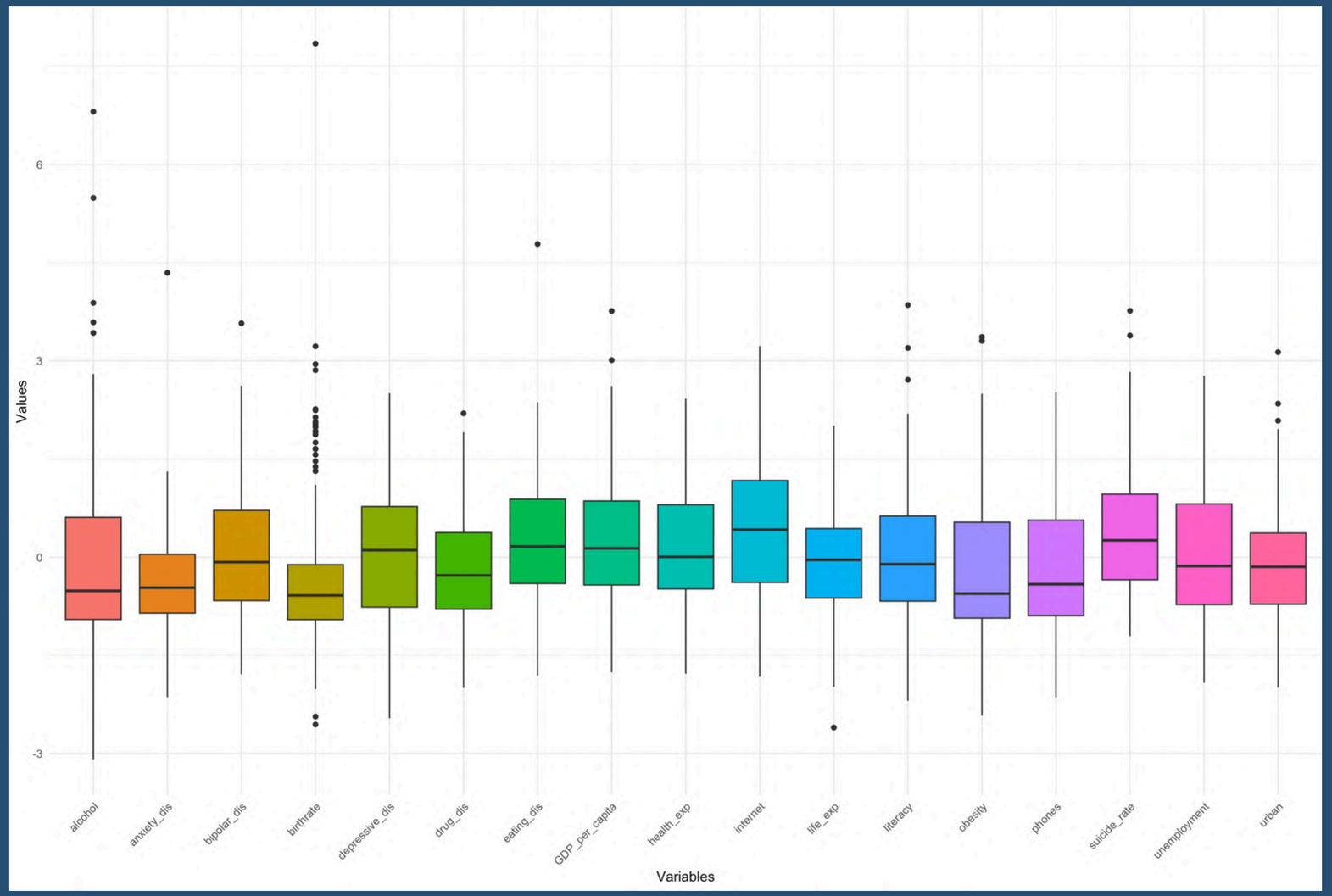
Literacy

Life expectancy

Birthrate

Exploratory Data Analysis (EDA)

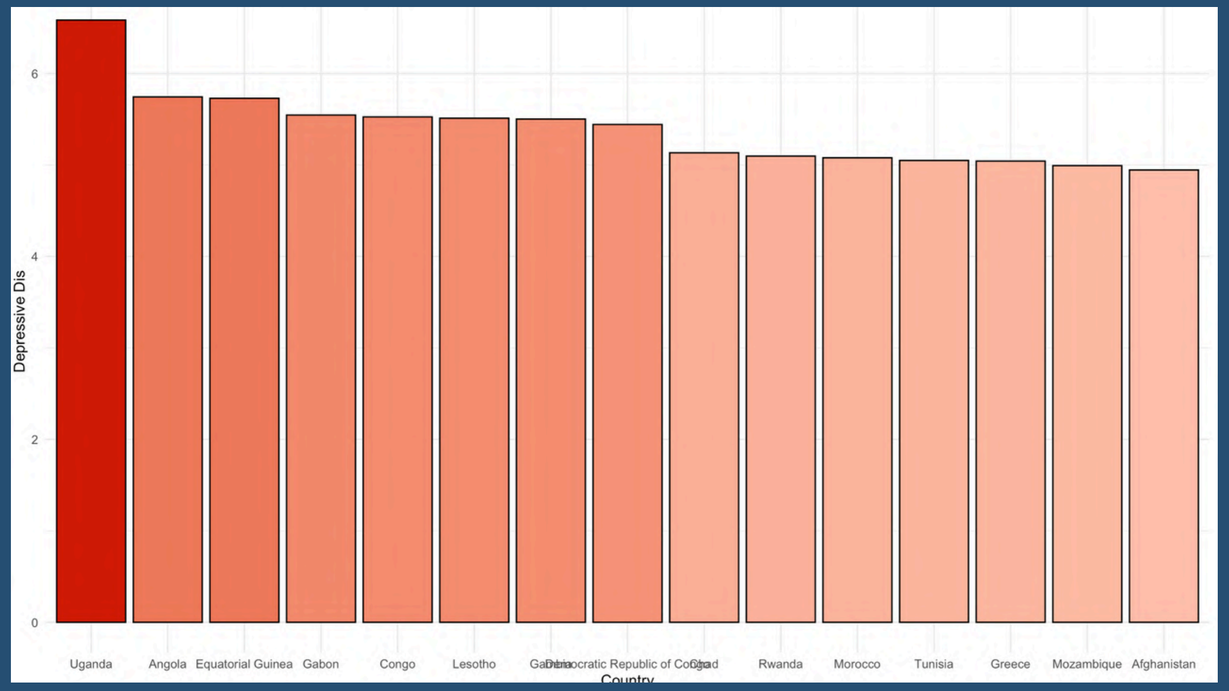
Distributions of the variables



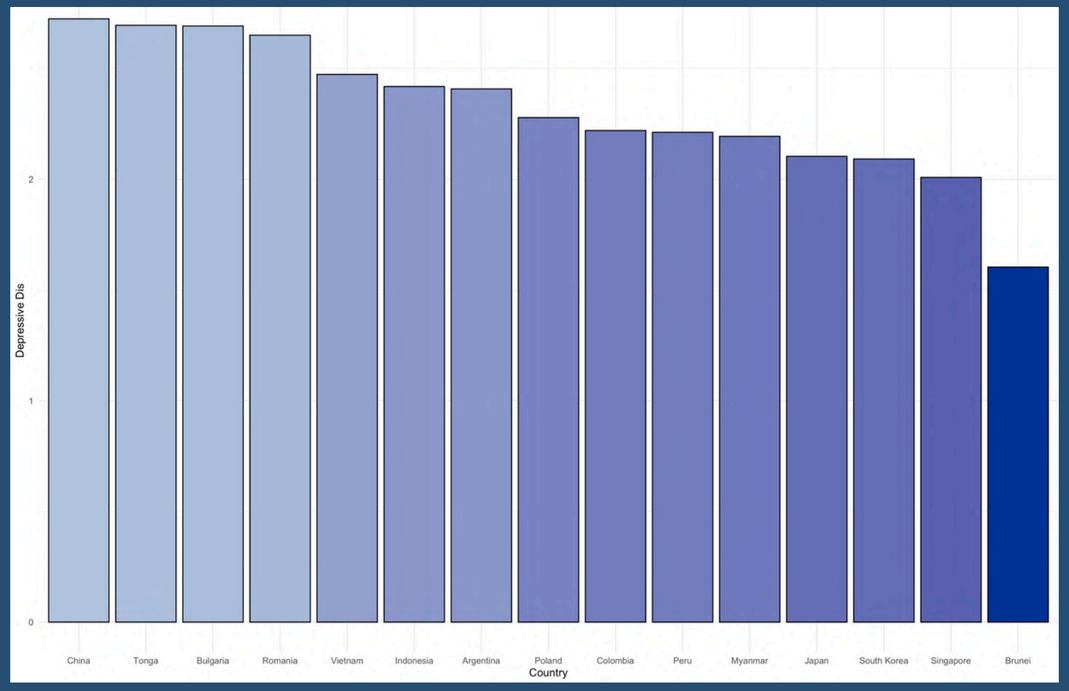
Outliers

- Alcohol
- Anxiety
- Bipolarity
- Birthrate
- Eating
- GDP per capita
- Life expectancy
- Obesity
- Suicide rate

Distribution of depression by country

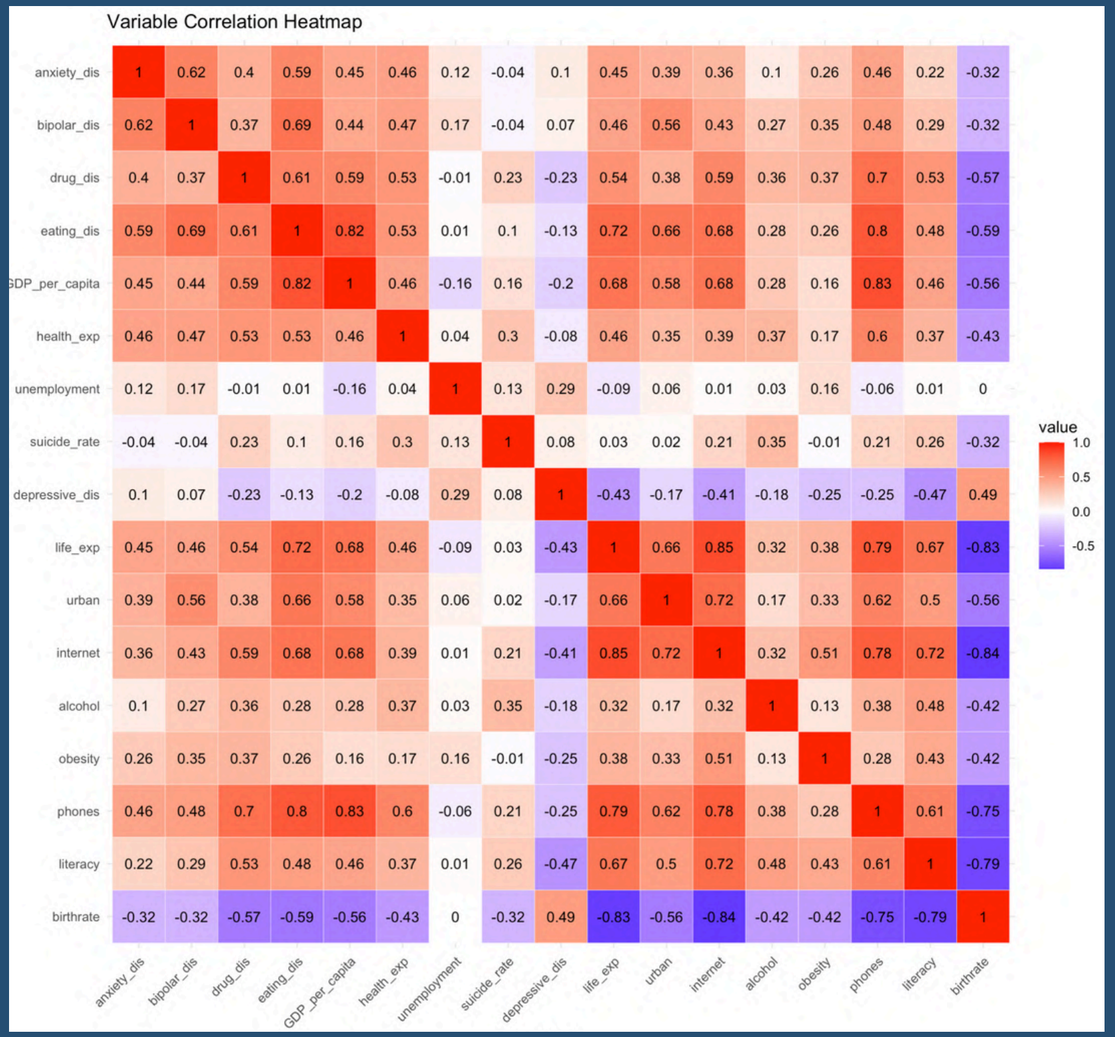


Most depressed countries



Less depressed countries

Correlation between Variables



Variables highly correlated with others

Eating disorders

GDP per capita

Internet

Unsupervised Learning

Principal Component Analysis

K-Means Clustering

Hierarchical Clustering

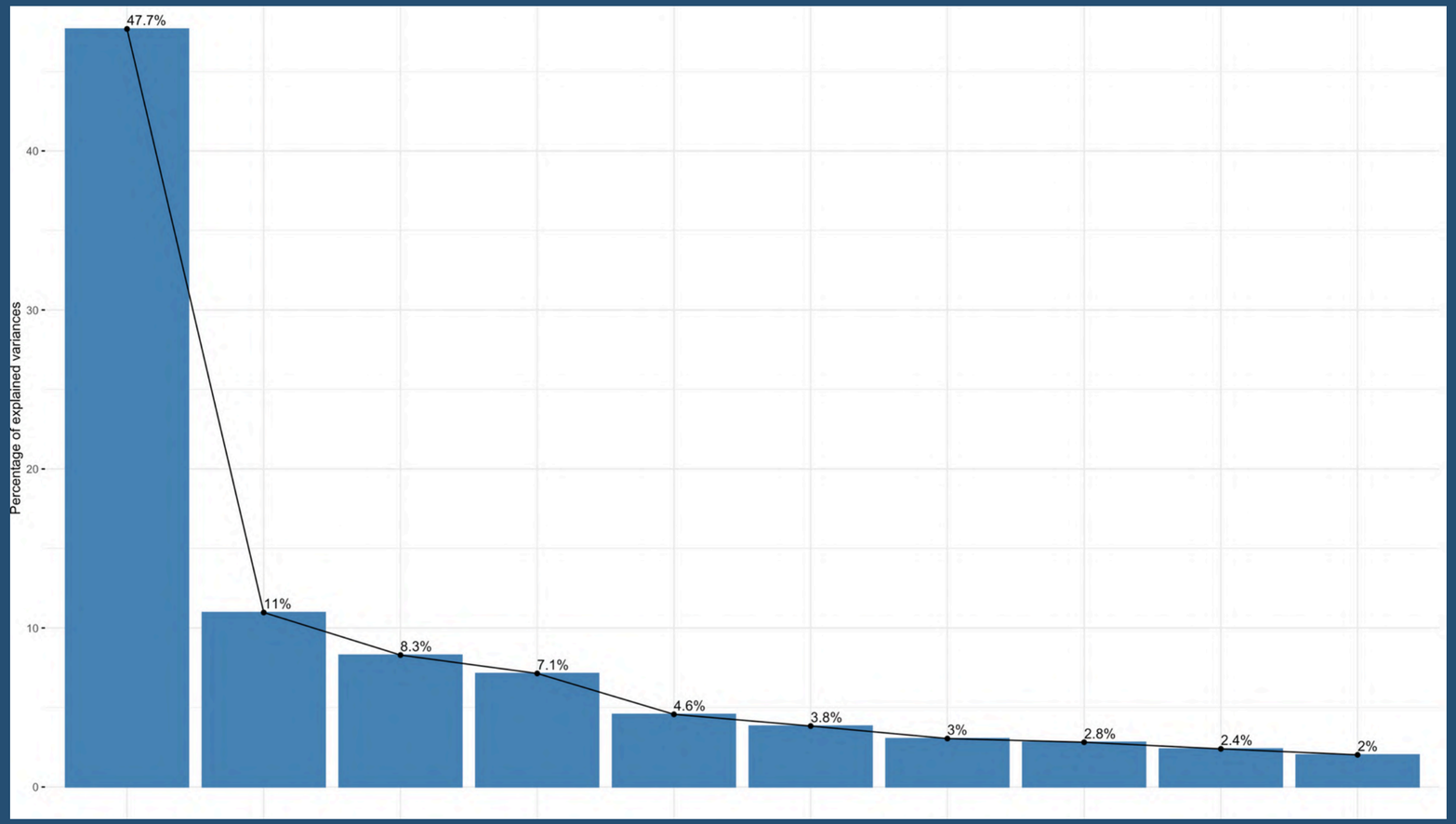
1. Principal component analysis (PCA)

- PCA reduces data complexity by determining the number of components to synthesise the data set.
- It generates principal components that efficiently represent the variances in the data.
- Normalise the data to remove differences in the scale of the variables.
- The number of components is chosen from the point of view of explaining the percentage of total variance.

Importance of components:

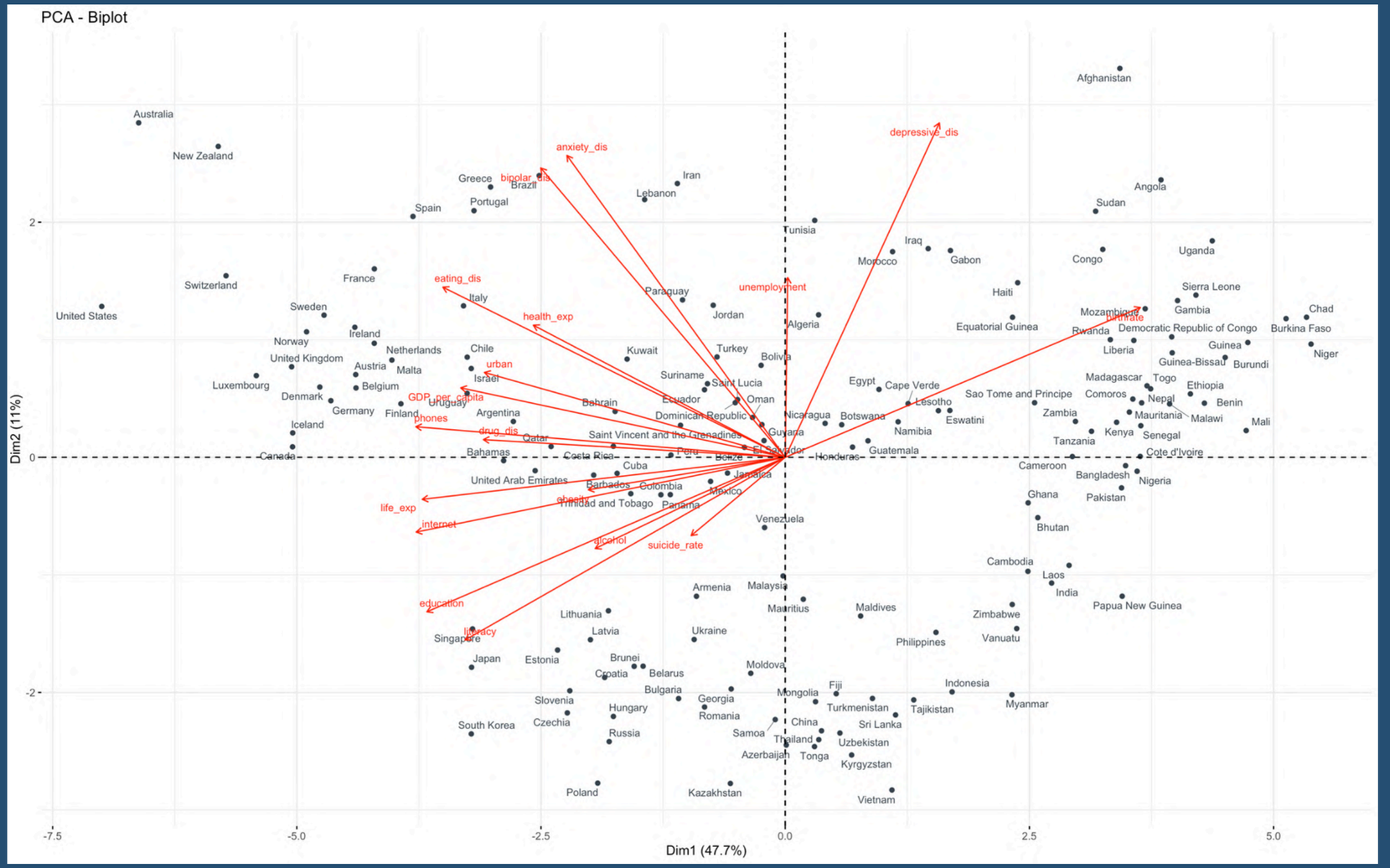
	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8
Standard deviation	2.9200499	1.4008534	1.21797092	1.12989232	0.90372177	0.8278564	0.73789146	0.70922003
Proportion of Variance	0.4767223	0.1097161	0.08293899	0.07137712	0.04566195	0.0383173	0.03044177	0.02812205
Cumulative Proportion	0.4767223	0.5864384	0.66937738	0.74075450	0.78641645	0.8247337	0.85517552	0.88329756

Scree Plot



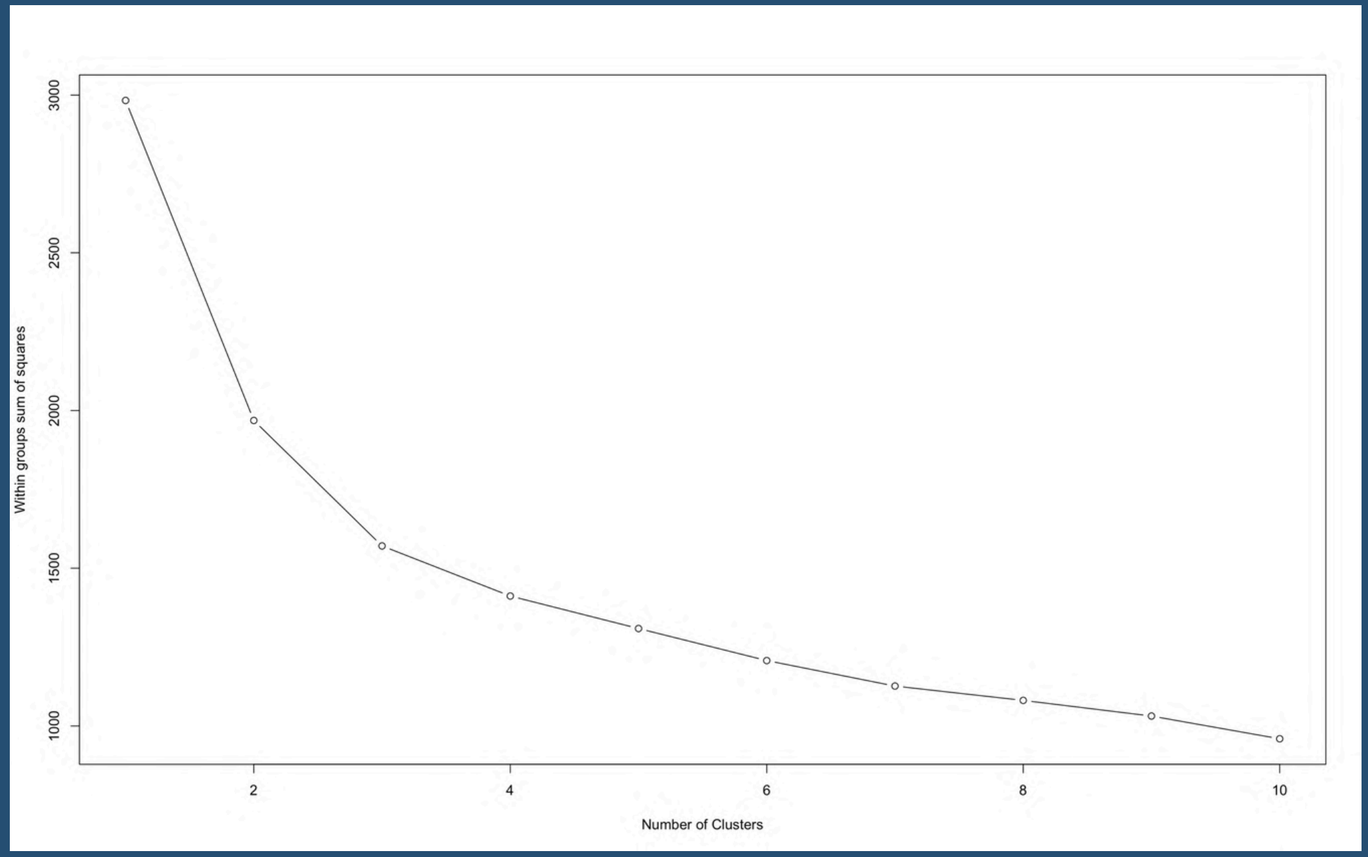
The first two components explain 58.64% of the total variance of the observations.

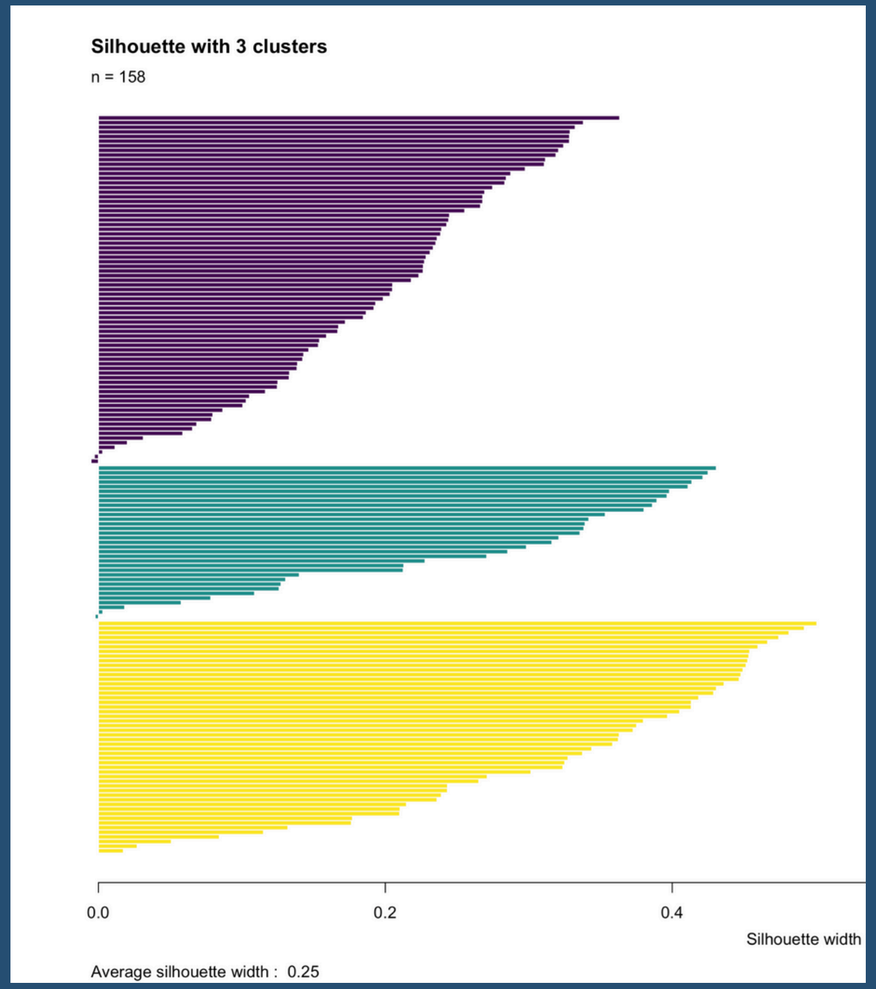
Biplot


PCA - Biplot
Dim2 (11%)
Dim1 (47.7%)
2
0
-2
-7.5
-5.0
-2.5
0.0
2.5
5.0
anxiety_dis
depressive_dis
bipolar_dis
eating_dis
unemployment
health_exp
urban
GDP_per_capita
phones
drug_dis
life_exp
internet
alcohol
suicide_rate
education
literacy
obesity
Afghanistan
Australia
New Zealand
Greece
Brazil
Iran
Lebanon
Spain
Portugal
Angola
Sudan
Tunisia
Iraq
Morocco
Gabon
Congo
Uganda
Switzerland
France
Haiti
Sierra Leone
United States
Sweden
Italy
Paraguay
Jordan
Algeria
Equatorial Guinea
Mozambique
Gambia
Chad
Ireland
Rwanda
Democratic Republic of Congo
Burkina Faso
Norway
Netherlands
Chile
Kuwait
Turkey
Bolivia
Liberia
Guinea
Niger
United Kingdom
Austria
Malta
Suriname
Saint Lucia
Guinea-Bissau
Burundi
Luxembourg
Israel
Egypt
Cape Verde
Madagascar
Togo
Ethiopia
Denmark
Belgium
Ecuador
Oman
Lesotho
Sao Tome and Principe
Comoros
Nepal
Benin
Germany
Finland
Uruguay
Bahrain
Dominican Republic
Nicaragua
Botswana
Eswatini
Zambia
Mauritania
Malawi
Mali
Iceland
Argentina
Namibia
Kenya
Senegal
Qatar
Saint Vincent and the Grenadines
Guyana
Tanzania
Canada
Bahamas
Peru
El Salvador
Guatemala
Cote d'Ivoire
Costa Rica
Belize
Honduras
Cameroon
Cuba
Bangladesh
United Arab Emirates
Jamaica
Nigeria
Barbados
Colombia
Mexico
Ghana
Pakistan
Trinidad and Tobago
Panama
Venezuela
Bhutan
Cambodia
Laos
Armenia
Malaysia
India
Mauritius
Maldives
Papua New Guinea
Lithuania
Zimbabwe
Singapore
Latvia
Ukraine
Philippines
Vanuatu
Japan
Estonia
Brunei
Moldova
Croatia
Belarus
Fiji
Indonesia
Slovenia
Bulgaria
Georgia
Mongolia
Turkmenistan
Tajikistan
Myanmar
Hungary
Romania
South Korea
Czechia
China
Sri Lanka
Russia
Samoa
Thailand
Uzbekistan
Azerbaijan
Tonga
Kyrgyzstan
Poland
Kazakhstan
Vietnam

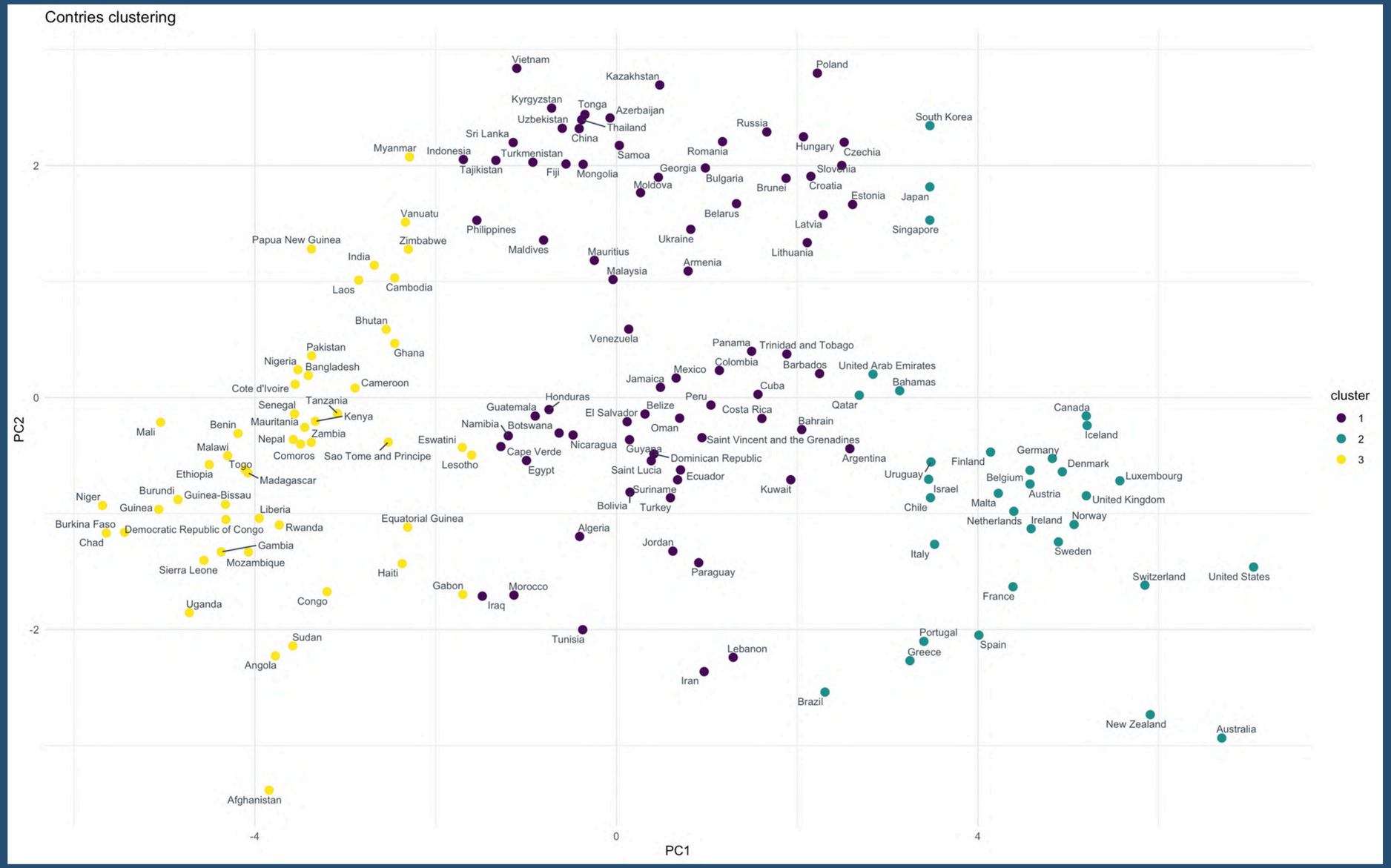
2. K-Means Clustering

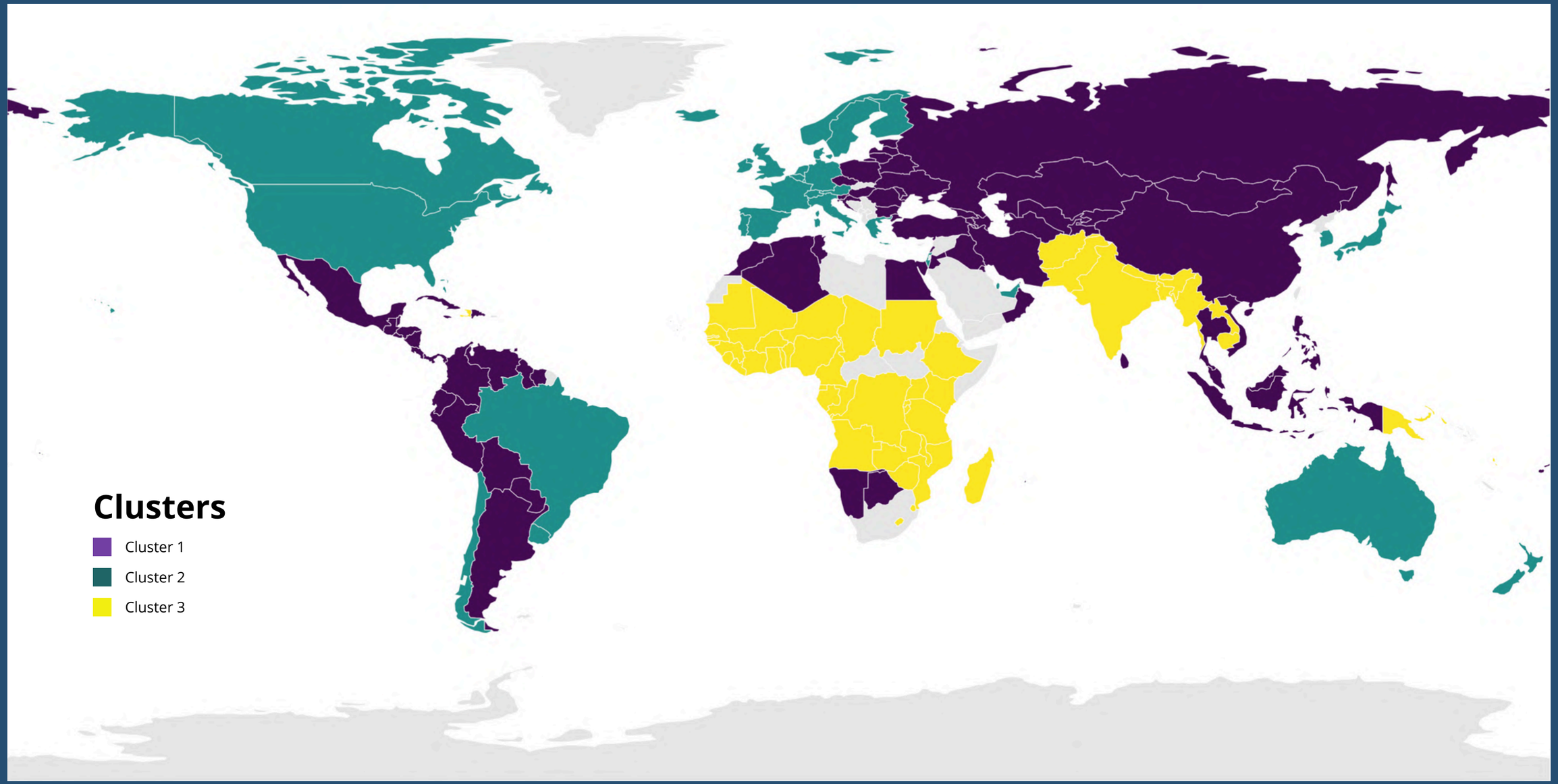
Calculate the optimal number of clusters with the Within Sum of Squares (WSS) plot, the Silhouette method and and the NbClust method (13 proposed 3 as the best number of clusters)

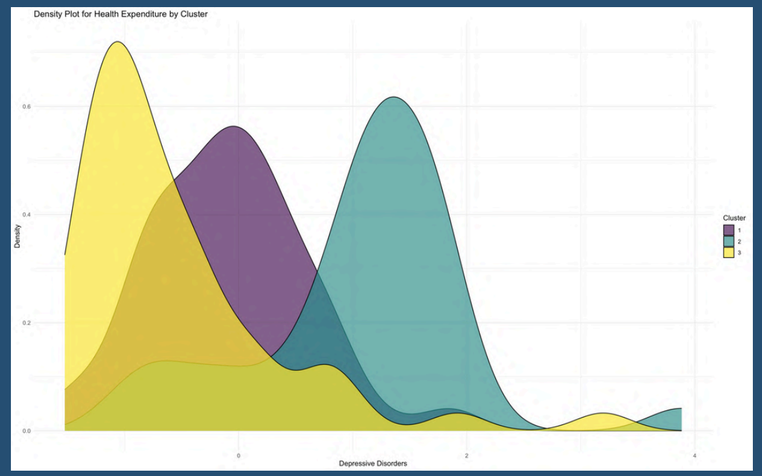




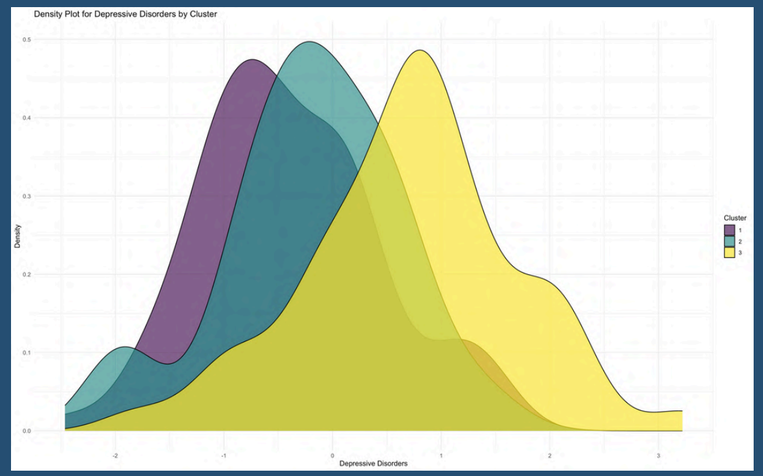

```
* Among all indices:
* 6 proposed 2 as the best number of clusters
* 13 proposed 3 as the best number of clusters
* 3 proposed 5 as the best number of clusters
* 1 proposed 6 as the best number of clusters

                 ***** Conclusion *****

* According to the majority rule, the best number of clusters is  3
```


Contries clustering
Vietnam
Poland
Kazakhstan
Kyrgyzstan
Tonga
Azerbaijan
Uzbekistan
Thailand
China
Russia
South Korea
Sri Lanka
Myanmar
Indonesia
Turkmenistan
Samoa
Romania
Hungary
Czechia
Tajikistan
Fiji
Mongolia
Georgia
Slovenia
Moldova
Bulgaria
Brunei
Croatia
Estonia
Japan
Belarus
Latvia
Singapore
Vanuatu
Philippines
Ukraine
Papua New Guinea
Zimbabwe
Maldives
Mauritius
Lithuania
India
Armenia
Malaysia
Laos
Cambodia
Bhutan
Venezuela
Panama
Trinidad and Tobago
Pakistan
Ghana
Nigeria
Bangladesh
Colombia
Barbados
United Arab Emirates
Mexico
Jamaica
Bahamas
Cameroon
Cote d'Ivoire
Cuba
Honduras
Peru
Qatar
Canada
Senegal
Tanzania
Guatemala
Belize
Costa Rica
Kenya
Mauritania
El Salvador
Bahrain
Iceland
Mali
Benin
Namibia
Botswana
Oman
Zambia
Nepal
Eswatini
Nicaragua
Saint Vincent and the Grenadines
Germany
Malawi
Comoros
Sao Tome and Principe
Cape Verde
Guyana
Argentina
Finland
Denmark
Togo
Lesotho
Egypt
Dominican Republic
Saint Lucia
Ecuador
Uruguay
Belgium
Luxembourg
Ethiopia
Madagascar
Kuwait
Israel
Austria
Niger
Burundi
Guinea-Bissau
Surinam
Malta
United Kingdom
Guinea
Liberia
Bolivia
Turkey
Chile
Netherlands
Ireland
Norway
Burkina Faso
Equatorial Guinea
Rwanda
Democratic Republic of Congo
Algeria
Chad
Jordan
Sweden
Gambia
Italy
Mozambique
Sierra Leone
Haiti
Paraguay
Switzerland
United States
Gabon
Morocco
France
Congo
Iraq
Uganda
Tunisia
Portugal
Spain
Sudan
Greece
Lebanon
Angola
Iran
Brazil
New Zealand
Australia
Afghanistan
PC2
PC1
2
0
-2
-4
0
4
cluster
1
2
3

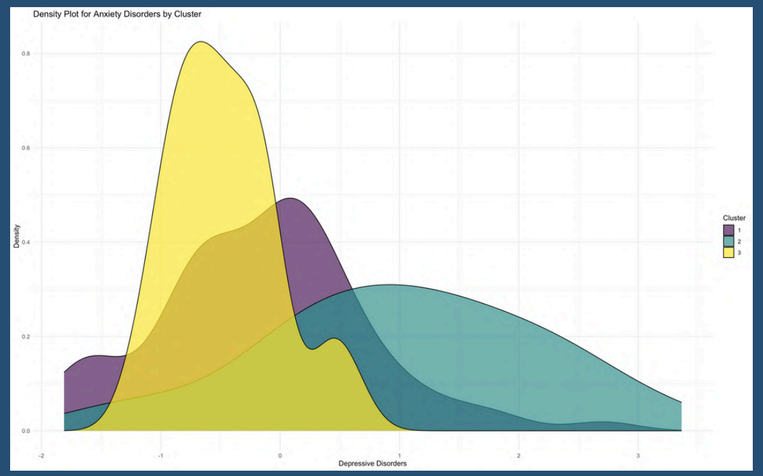

Clusters
Cluster 1
Cluster 2
Cluster 3



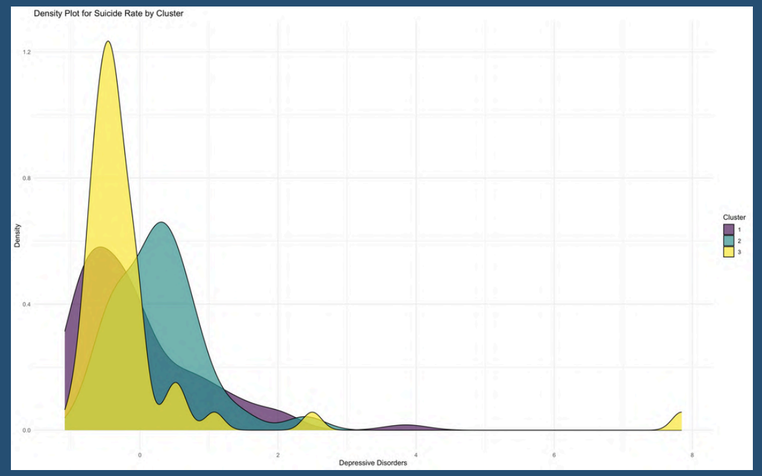
Health expenditure



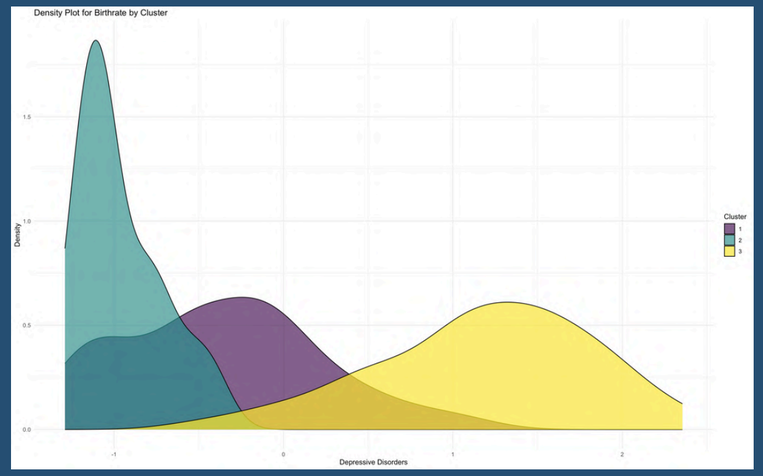
Depressive disorders



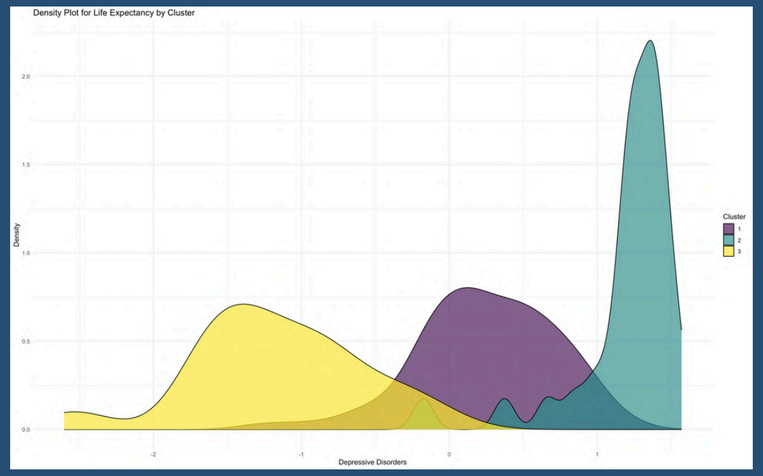
Anxiety disorders



Suicide rate



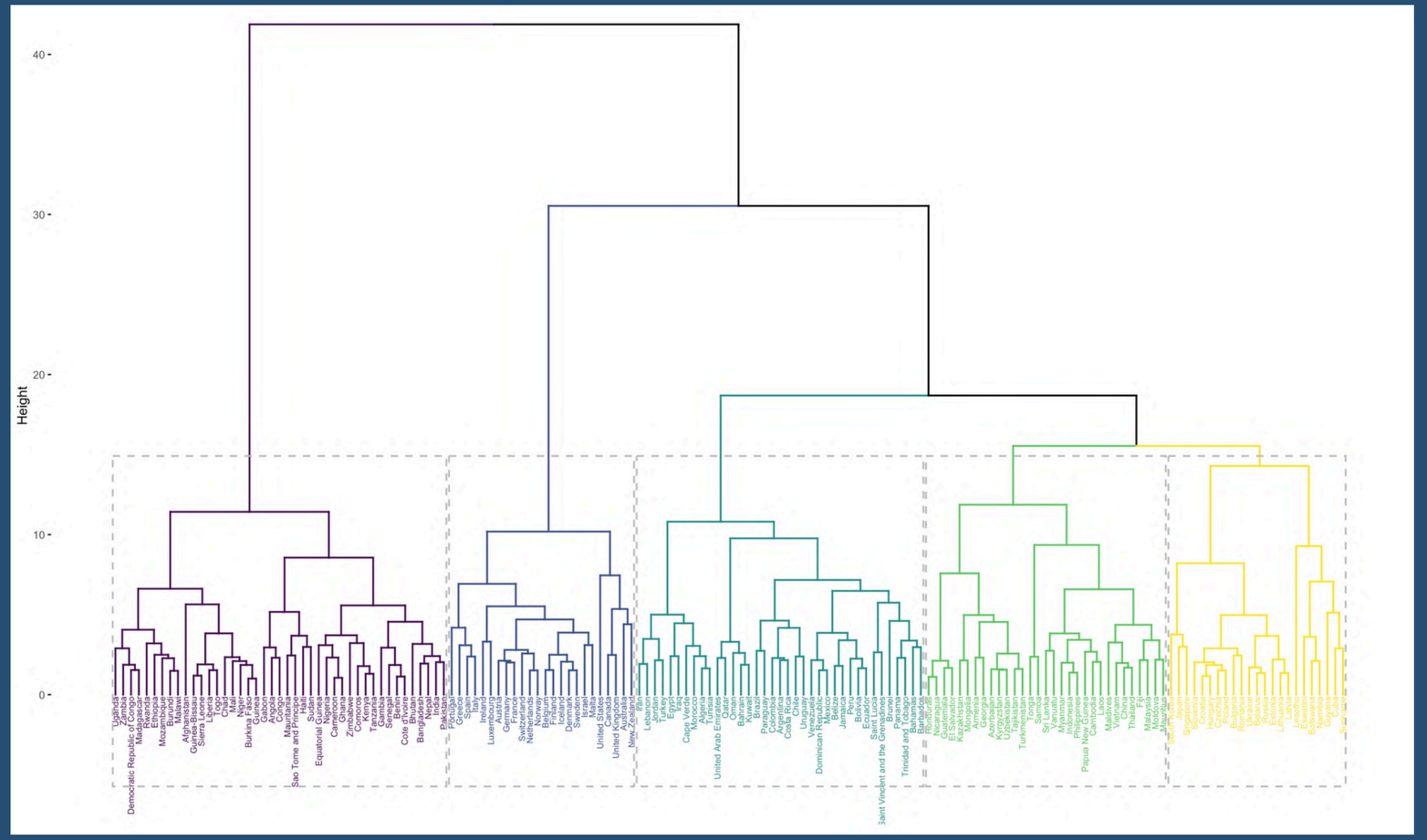
Birthrate

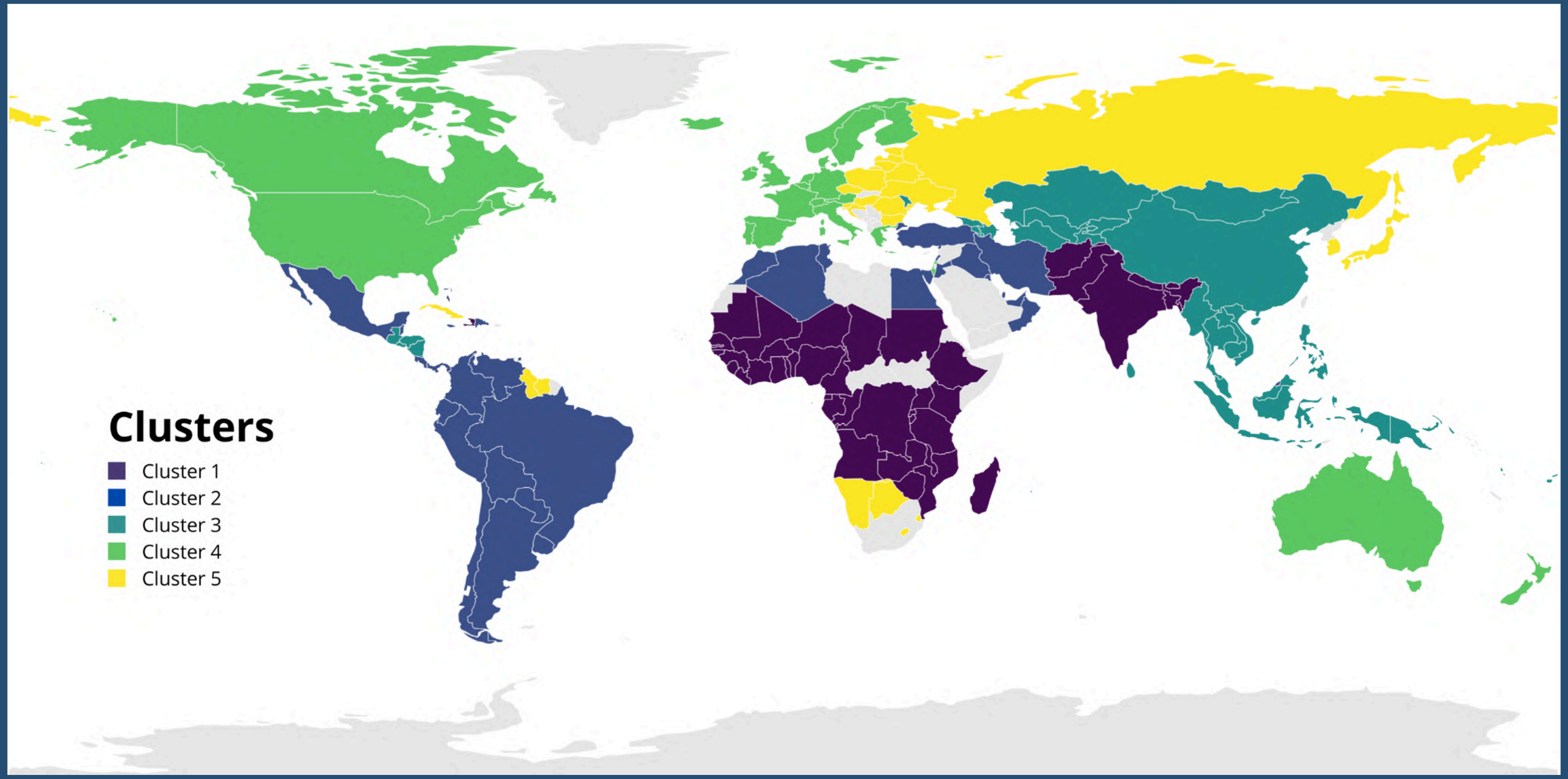


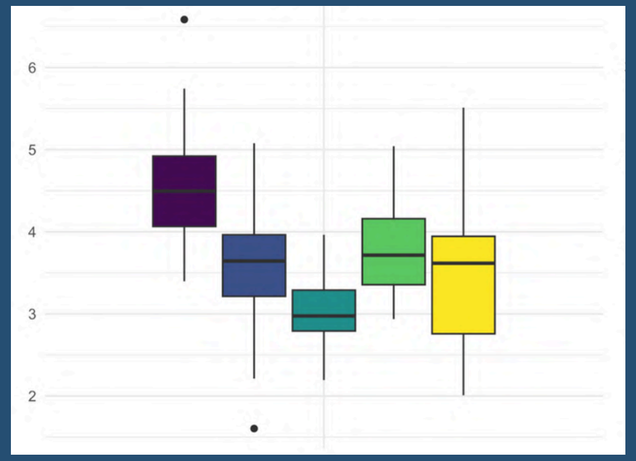
Life expectancy

3. Hierarchical Clustering

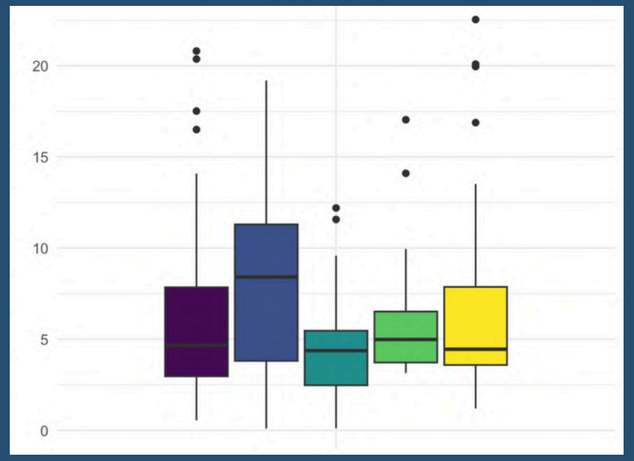
For a better visualization the Ward’s method was used.



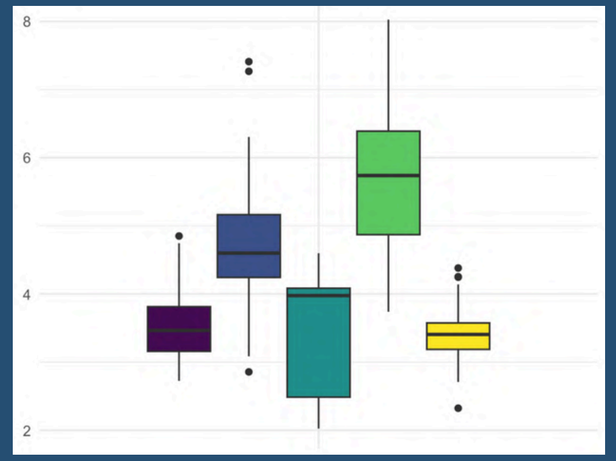

Clusters
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5



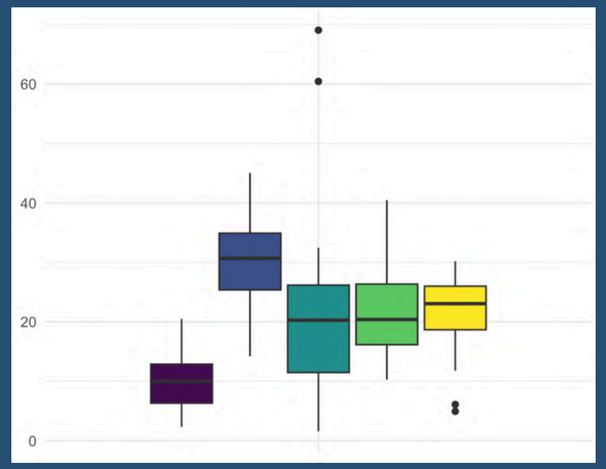
Depressive disorders



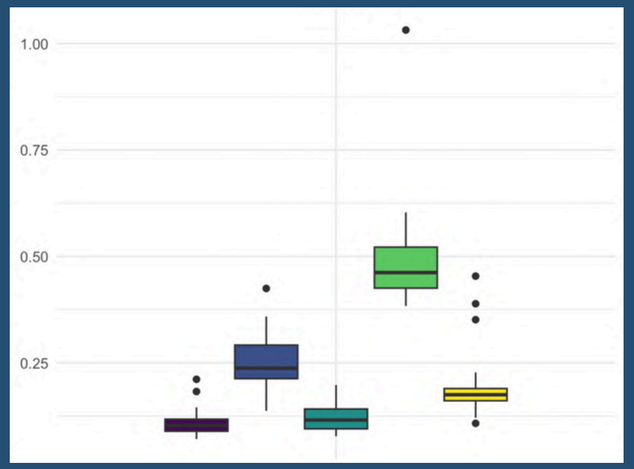
Unemployment



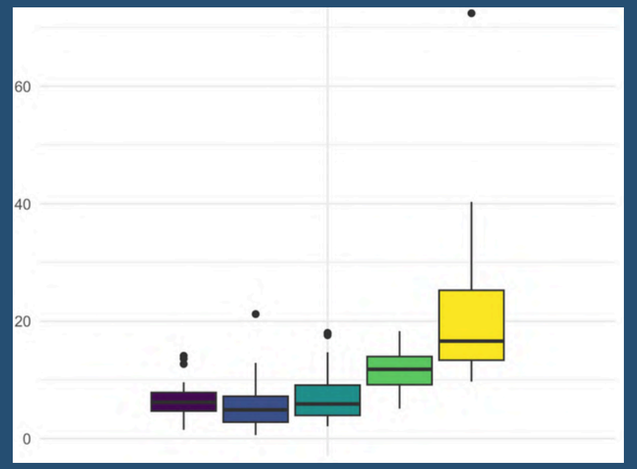
Anxiety disorders



Obesity



Eating disorders



Suicide rate

Supervised Learning

Assumptions

Linear and stepwise regression

Ridge and Lasso regression

Decision tree

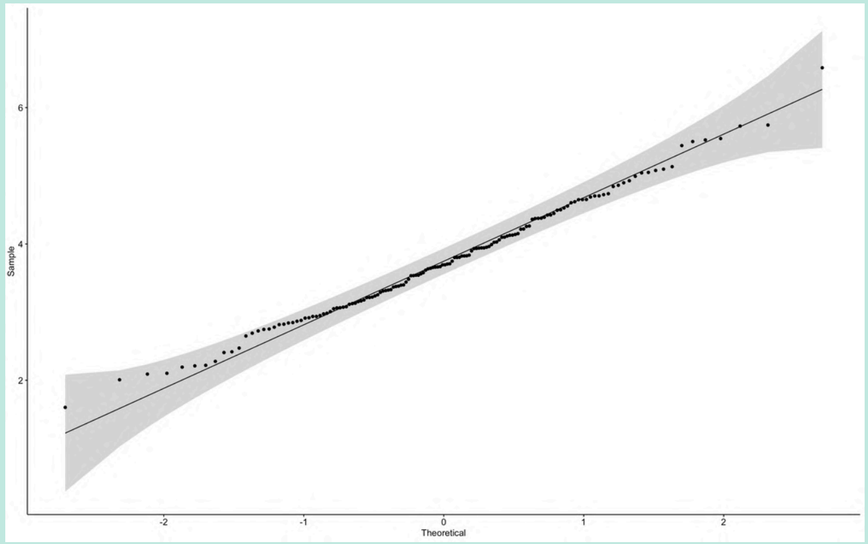
1. Assumptions

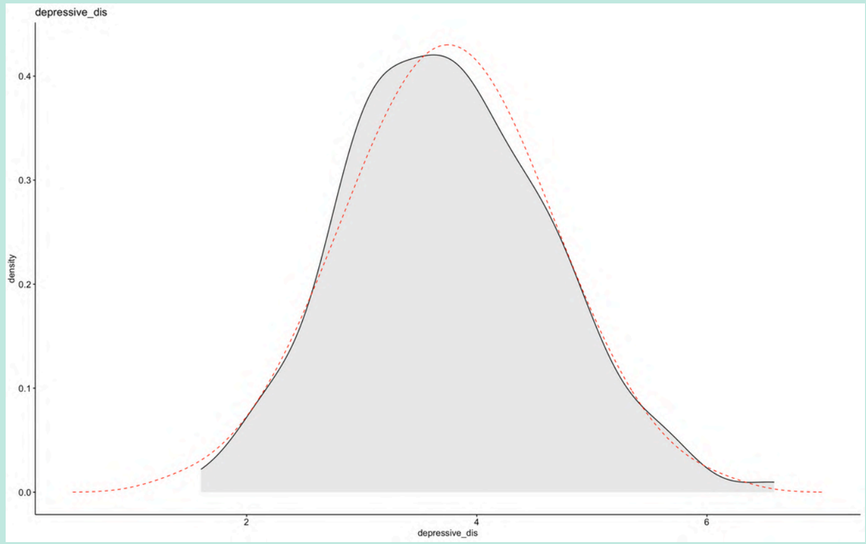
1. Remove outliers

```
> states_to_remove <- c("Mongolia", "El Salvador", "Portugal", "New Zealand", "Niger", "Australia",
+                       "Luxembourg", "Nigeria", "Tonga", "Samoa", "Lesotho", "USA", "Afghanistan")
```

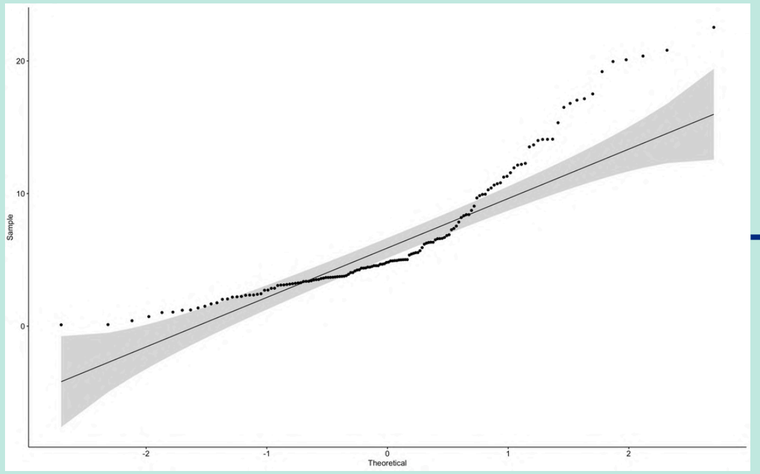
2. Homogeneity of the variables → If they follow a normal distribution (eating disorder, GDP per capita, unemployment, suicide rate, literacy)

Depressive disorders

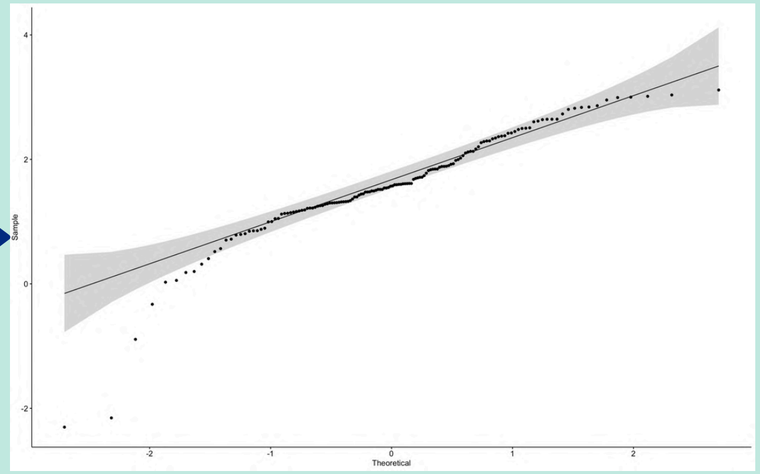




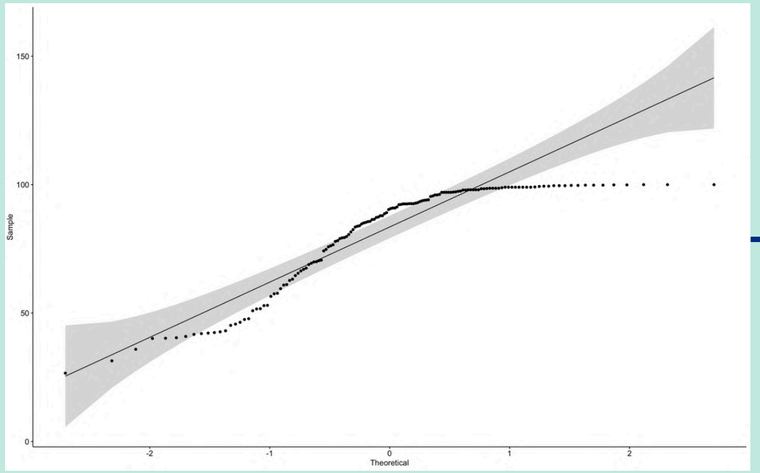
Unemployment



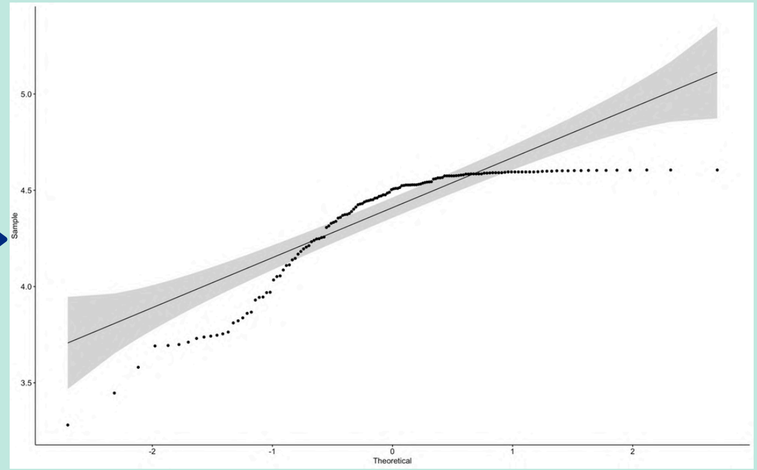
log(Unemployment)



Literacy



log(Literacy)



3. Eliminate variables with a very high multicollinearity value by calculating the VIF (Variance Inflation Factor)

```
> vif(model)
                        GVIF Df GVIF^(1/(2*Df))
anxiety_dis         1.991716  1        1.411282
bipolar_dis         4.350859  1        2.085871
drug_dis            2.450013  1        1.565252
log(eating_dis)    15.074555  1        3.882596
log(GDP_per_capita) 19.397554  1        4.404265
health_exp          2.643688  1        1.625942
income             10.183444  2        1.786379
log(unemployment)   1.313563  1        1.146108
log(suicide_rate)   1.781995  1        1.334914
life_exp            7.582576  1        2.753648
urban               3.232048  1        1.797790
internet            9.899875  1        3.146407
alcohol             1.766365  1        1.329047
education           5.340972  1        2.311054
obesity             2.430976  1        1.559159
phones              7.599228  1        2.756670
birthrate           6.316713  1        2.513307
```

Variables will not be used for the final model:

log(eating_dis)

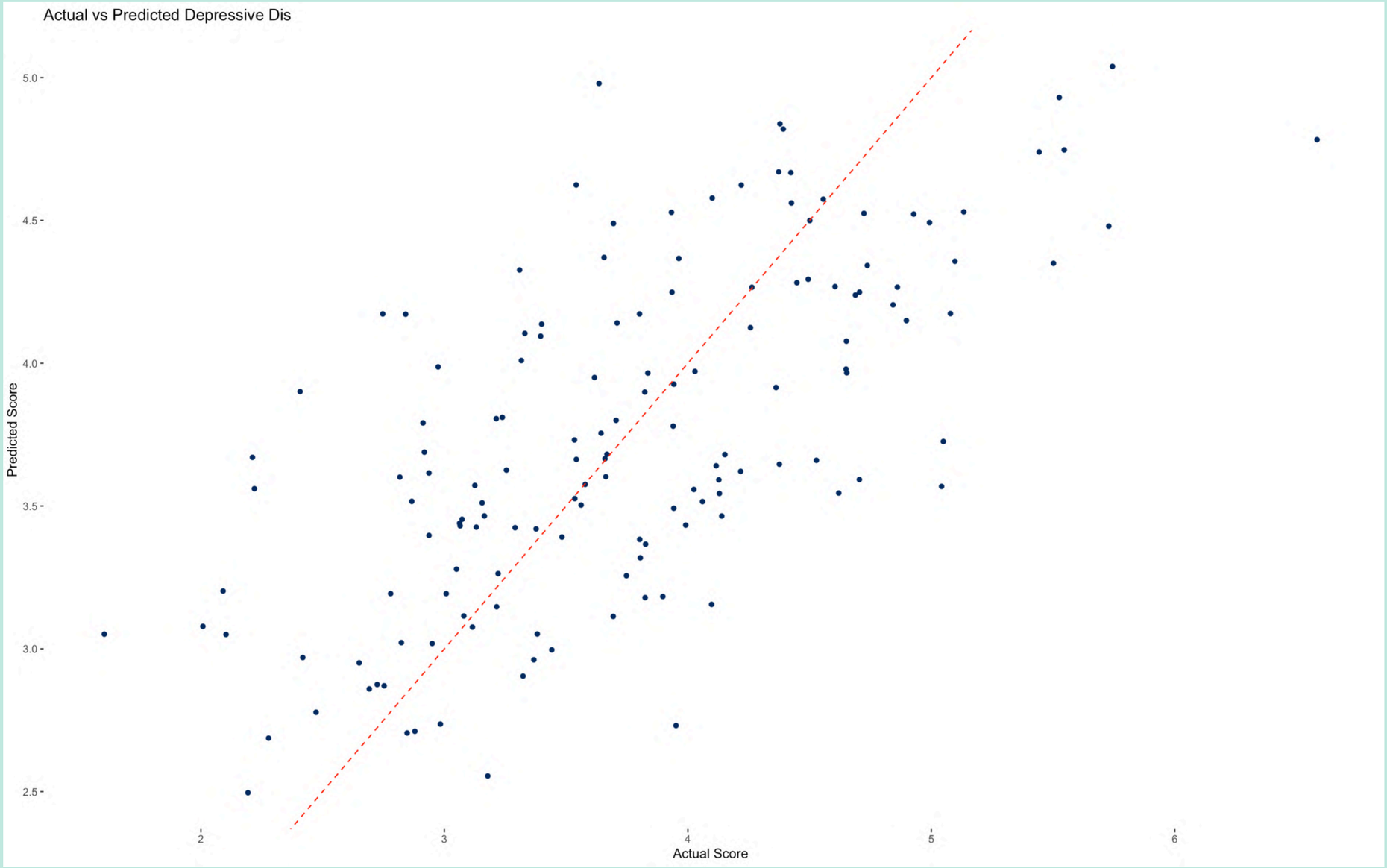
log(GDP_per_capita)

internet

2. Linear Regression

```
Call:
lm(formula = depressive_dis ~ anxiety_dis + bipolar_dis + drug_dis +
    health_exp + income + log(unemployment) + log(suicide_rate) +
    life_exp + urban + alcohol + education + obesity + phones +
    birthrate, data = mental_health)

Residuals:
     Min       1Q   Median       3Q      Max
-1.60896 -0.38587 -0.01685  0.42843  1.67778

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       2.1880066  1.6042142   1.364 0.174952
anxiety_dis       0.1806155  0.0694197   2.602 0.010349 *
bipolar_dis       0.5780356  0.4238270   1.364 0.174972
drug_dis         -0.1757608  0.2306290  -0.762 0.447385
health_exp       -0.0500836  0.0335517  -1.493 0.137931
incomeLow         0.4547990  0.3507703   1.297 0.197075
incomeMiddle      0.0663187  0.2179871   0.304 0.761437
log(unemployment) 0.2030696  0.0725087   2.801 0.005879 **
log(suicide_rate) 0.3723241  0.1017822   3.658 0.000368 ***
life_exp         -0.0163454  0.0187516  -0.872 0.384990
urban             0.0021438  0.0039381   0.544 0.587119
alcohol          -0.1182609  0.0905167  -1.307 0.193686
education        -0.0220145  0.0392631  -0.561 0.575972
obesity           0.0011141  0.0078990   0.141 0.888055
phones            0.0011674  0.0007322   1.594 0.113290
birthrate         0.0389696  0.0119459   3.262 0.001412 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6661 on 130 degrees of freedom
Multiple R-squared:  0.4876,    Adjusted R-squared:  0.4284
F-statistic: 8.246 on 15 and 130 DF,  p-value: 6.527e-13
```

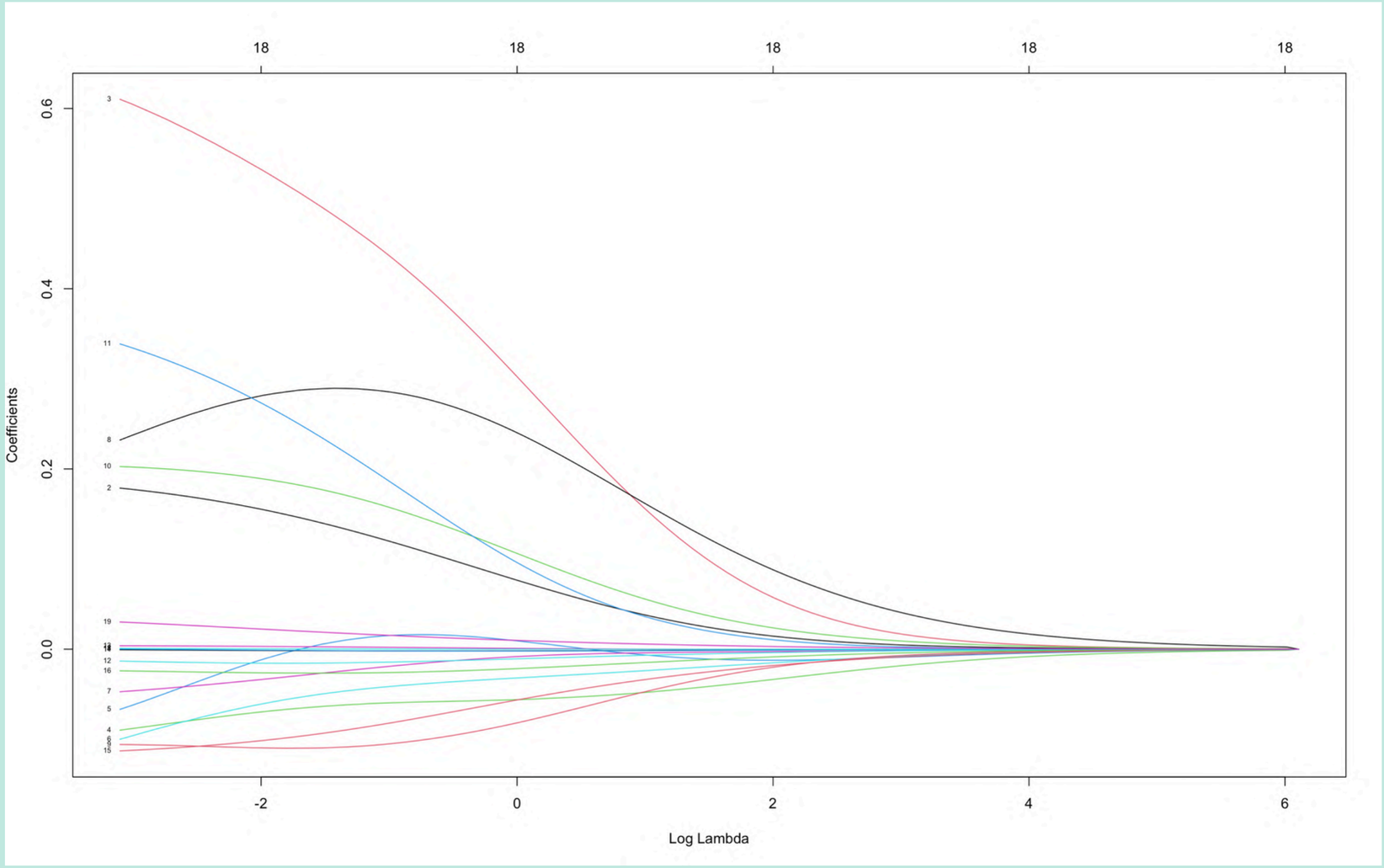
```
Call:
lm(formula = depressive_dis ~ anxiety_dis + bipolar_dis + log(unemployment) +
    log(suicide_rate) + alcohol + birthrate, data = mental_health)

Residuals:
     Min       1Q   Median       3Q      Max
-1.49324 -0.40390 -0.00953  0.48359  1.80194

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        0.605659   0.403492   1.501  0.13561
anxiety_dis        0.161878   0.064428   2.513  0.01313 *
bipolar_dis        0.620896   0.352225   1.763  0.08013 .
log(unemployment)  0.201832   0.067751   2.979  0.00341 **
log(suicide_rate)  0.407043   0.090109   4.517 1.33e-05 ***
alcohol           -0.164283   0.081683  -2.011  0.04623 *
birthrate          0.051602   0.005766   8.950 2.03e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6665 on 139 degrees of freedom
Multiple R-squared:  0.4515,    Adjusted R-squared:  0.4279
F-statistic: 19.07 on 6 and 139 DF,  p-value: 3.949e-16
```


Actual vs Predicted Depressive Dis
Predicted Score
5.0
4.5
4.0
3.5
3.0
2.5
2
3
4
5
6
Actual Score

3. Stepwise Regression

```
Call:
lm(formula = depressive_dis ~ anxiety_dis + bipolar_dis + health_exp +
    income + log(unemployment) + log(suicide_rate) + alcohol +
    birthrate, data = mental_health)

Residuals:
     Min       1Q   Median       3Q      Max
-1.63698 -0.42347 -0.01655  0.47202  1.58659

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        0.971848   0.436889   2.224  0.02777 *
anxiety_dis        0.185145   0.065983   2.806  0.00575 **
bipolar_dis        0.638983   0.370214   1.726  0.08662 .
health_exp        -0.045239   0.030562  -1.480  0.14113
incomeLow          0.422145   0.289182   1.460  0.14665
incomeMiddle      -0.067068   0.163640  -0.410  0.68256
log(unemployment)  0.228189   0.070007   3.260  0.00141 **
log(suicide_rate)  0.396714   0.092391   4.294 3.32e-05 ***
alcohol           -0.144112   0.082917  -1.738  0.08447 .
birthrate          0.040246   0.008461   4.757 4.95e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6598 on 136 degrees of freedom
Multiple R-squared:  0.474,     Adjusted R-squared:  0.4392
F-statistic: 13.62 on 9 and 136 DF,  p-value: 2.054e-15
```

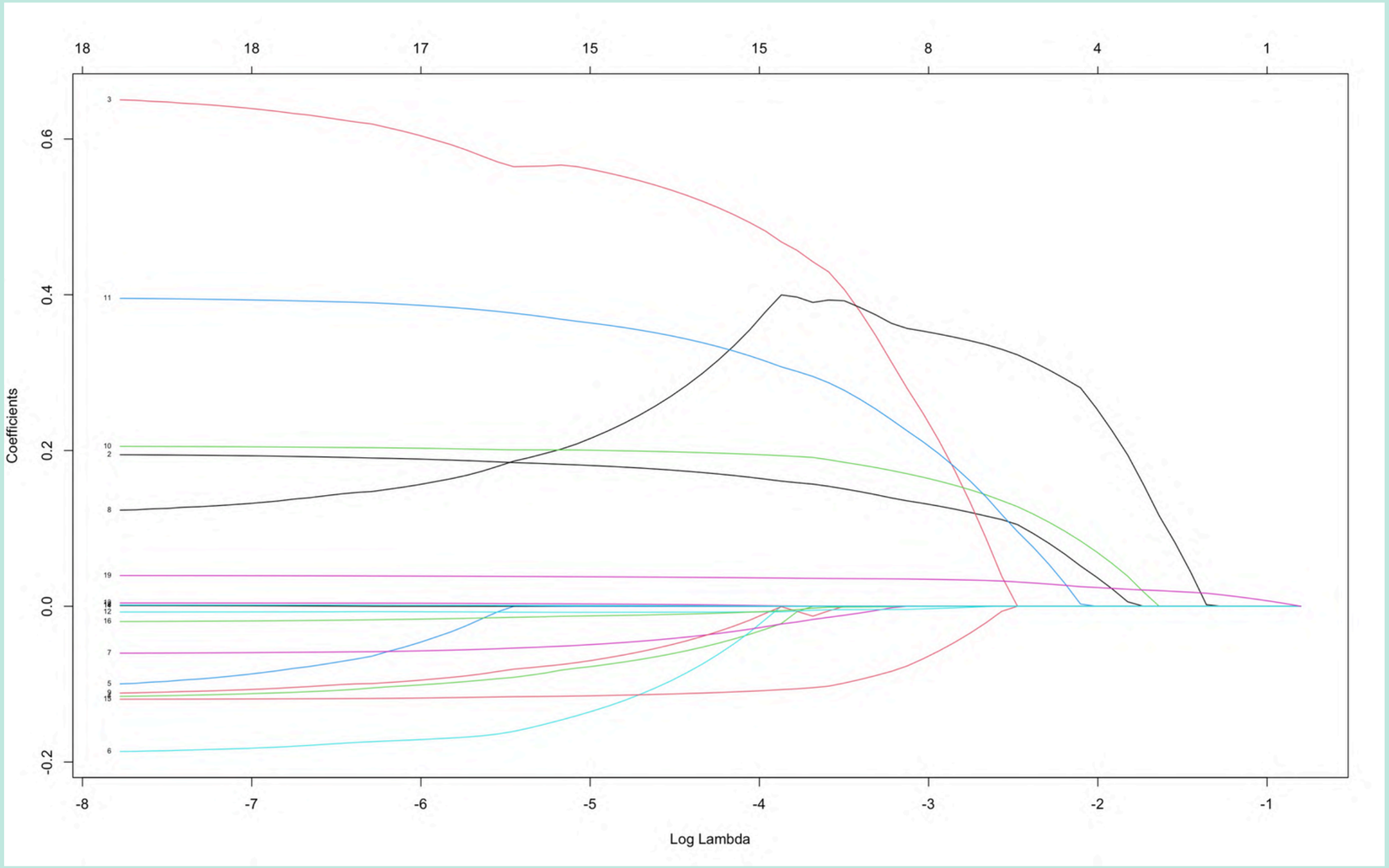
4. Ridge Regression



	s0
(Intercept)	3.6523698441
(Intercept)	.
anxiety_dis	0.1490839699
bipolar_dis	0.5160296980
drug_dis	-0.0658660369
log(eating_dis)	-0.0009687116
log(GDP_per_capita)	-0.0552015280
health_exp	-0.0307686062
incomeLow	0.2868666502
incomeMiddle	-0.1089282411
log(unemployment)	0.1845678607
log(suicide_rate)	0.2575472610
life_exp	-0.0156982250
urban	0.0029036475
internet	-0.0016320239
alcohol	-0.0987266355
education	-0.0266074027
obesity	-0.0012332192
phones	0.0004010886
birthrate	0.0207745409

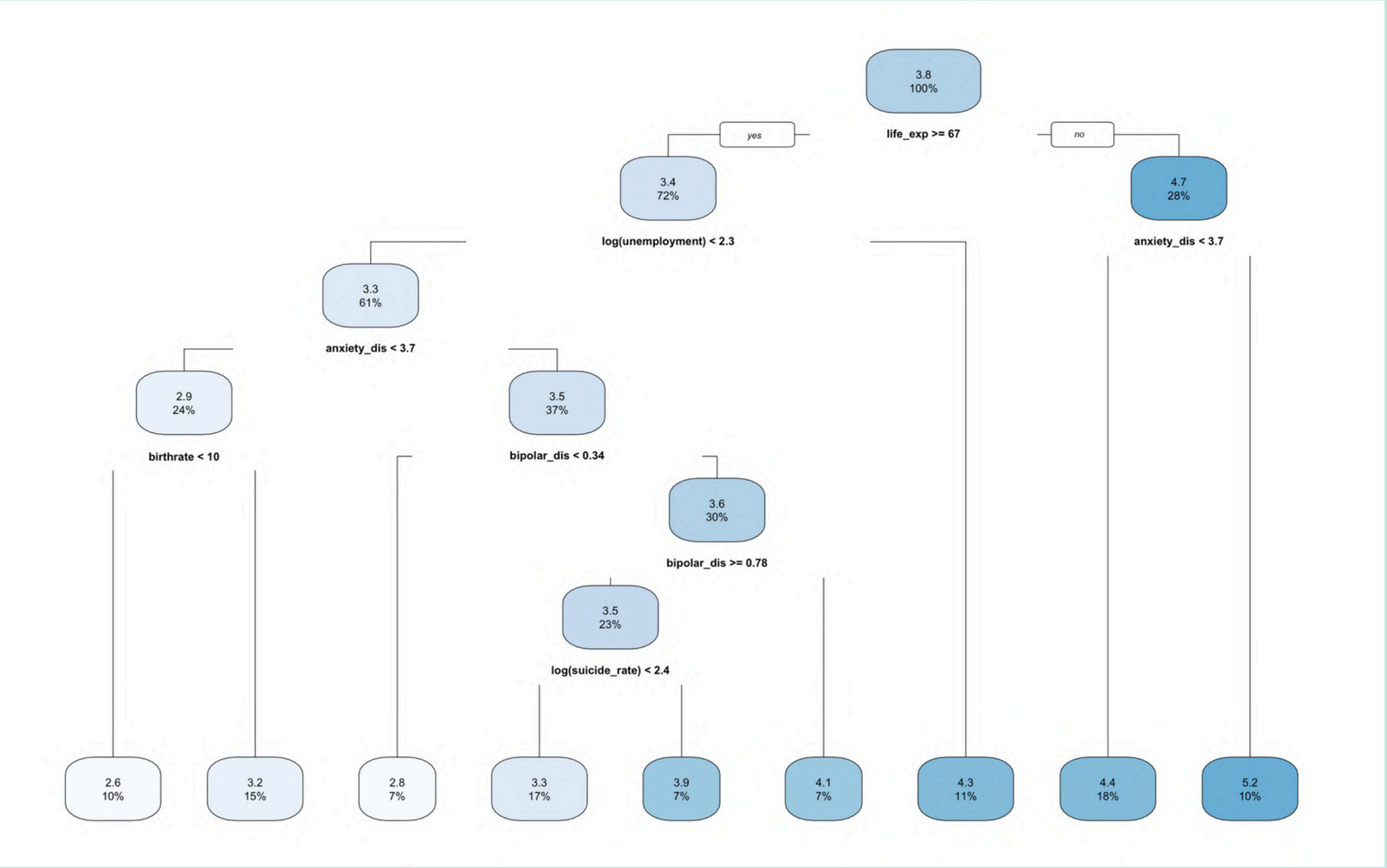
5. Lasso Regression

	s0
(Intercept)	1.7763498454
(Intercept)	.
anxiety_dis	0.1434526808
bipolar_dis	0.3475944184
drug_dis	.
log(eating_dis)	.
log(GDP_per_capita)	.
health_exp	-0.0053713106
incomeLow	0.3738072838
incomeMiddle	.
log(unemployment)	0.1780172829
log(suicide_rate)	0.2529490897
life_exp	-0.0043530475
urban	.
internet	.
alcohol	-0.0892071165
education	-0.0007081653
obesity	.
phones	.
birthrate	0.0352618319



	Ridge Regression	**Lasso Regression**
Mean Square Error (MSE)	**0.4084427**	0.4221251
R-squared value	**0.4763796**	0.4583483

6. Decision Tree



Mean Square Error (MSE)	0.270865
R-squared value	0.6933

CONCLUSIONS

The study reveals that mental health problems are not exclusive to less developed countries; on the contrary, richer and more advanced countries are greatly affected.

Unsupervised clustering techniques reveal two main clusters of nations. The first cluster includes economically developed countries with high mental health disorder rates, challenging the notion that economic prosperity guarantees good mental health. The second cluster comprises nations from Central Africa and South-East Asia, facing a high incidence of depression and related challenges.

Key variables significantly influencing depression rates, encompassing social, economic, and cultural dimensions, are identified. These include anxiety, bipolar disorder, unemployment, suicide rates, alcohol consumption, birth rate, public health investment, education level and life expectancy.