A Worldwide Exploration of Depression: Understanding the Factors

Tommaso Premoli - Nr. 34221A

2024

Abstract This project aims to comprehensively examine the determinants of mental health and depression index at a global level through statistical learning methods. Through a comprehensive analysis of the socio-economic, demographic and health characteristics of countries, the objective is to identify common patterns and distinct geographical clusters based on the similarity of individual observations. Another purpose of the research is to develop a predictive model to estimate the level of the population affected by depression in each country in order to identify priority areas for intervention and improve mental well-being globally. The study is done using a wide range of data taking into account various variables such as depression, GDP per capita, unemployment, current health expenditure and other relevant indicators, to provide an in-depth understanding of mental health worldwide and guide targeted policies and interventions.

Objective of the analysis Are there distinct patterns or clusters among countries based on socio-economic, demographic and health characteristics that can be associated with the mental health index? This research question focuses on exploring whether there are distinct patterns or clusters of countries that emerge when their socio-economic, demographic and health characteristics are considered. The goal is to identify whether there are associations between these variables and the mental health index of countries. For example, clusters of countries with similar characteristics could emerge that are associated with higher or lower levels of mental health.

What are the main socio-economic, demographic and health factors influencing the mental health index in different countries? This question wants to identify the key factors among socio-economic, demographic and health characteristics that have a significant impact on the mental health index in different countries. The aim is to identify the main drivers that influence mental health globally and to understand how these factors may vary between countries with different contexts.

Contents

1 Introduction

Attention to mental health has grown significantly in recent years, highlighting the importance of raising awareness of mental health issues. Ensuring access to psychological care is crucial to promote an optimal balance between mind and body for as many individuals as possible. The concept of mental health extends beyond individual well-being, embracing the concept of global wellness, which includes the care of both body and mind. Among the most common symptoms that threaten this balance are depression, anxiety, stress and eating disorders.

This study focuses on analysing depression, one of the most widespread mental illnesses, in order to understand the social, economic and psychological factors that influence its prevalence on a global scale. The observations and data cover the year 2019. The dependent variable used is the percentage of the population affected by depression in each country, thus allowing for the exploration of the influences of different socio-cultural contexts. By using these variables as predictor factors, the aim is not only to understand the determinants of depression, but also to predict and assess the likelihood of developing this disorder in a specific country context.

This research aims to identify priority areas for intervention to improve mental health globally, thus making a significant contribution to promoting the psychological well-being of the world's population.

In this research, we will adopt Statistical Learning methodologies to explore the collected observations in depth. The work will be divided into two main procedures: the unsupervised learning phase and the supervised learning phase. In the supervised learning phase, we will conduct an analysis using two specific methods: Principal Component Analysis (PCA) and Clustering. Through the application of PCA, we will aim to identify the most relevant variables and gain a deeper understanding of the factors affecting the spread of depression on a global scale. At the same time, through clustering, it will be possible to identify similar patterns and trends in the data, making it possible to identify groups of countries with common characteristics and specific needs regarding the mental well-being of the population on a global scale. Subsequently, in the supervised learning phase, we will develop predictive models in order to estimate the percentage of people with depression in a given nation. These models will take into account the complex relationships between the variables considered, enabling a more precise assessment of the situation and providing useful information for the formulation of targeted interventions. Regression techniques such as stepwise, Ridge and Lasso and the decision tree will be implemented.

At the conclusion of the research, it will be possible to determine whether economic variables, such as GDP per capita and the unemployment rate, together with government investment in the health sector and the average state income, or social variables, such as the percentage of the population suffering from bipolar disorders, anxiety, eating disorders and drug addiction, together with the suicide rate, can explain the number of people suffering from depression in a given nation.

1.1 Dataset and Variables

The final dataset used for this research consists of 158 countries and was created by merging data from different sources, all relating to the year 2019.

For a more detailed explanation, the meaning of each variable is given below:

- Depression (depressive dis) = Percentage of people with depression in that country during the year 2019.
- Anxiety disorder (anxiety dis) = Estimated share of people with anxiety disorders, whether or not they are diagnosed, based on representative surveys, medical data and statistical modelling.
- Bipolar disorder (bipolar dis) = Estimated share of males versus females who had bipolar disorder in the past year, whether or not they were diagnosed, based on representative surveys, medical data and statistical modelling.
- People with drug use disorders (drug dis) = Percentage of the total number of people with a drug use disorder divided by the number of inhabitants of the country. The International Classification of Diseases defines drug dependence as the presence of three or more indicators of dependence for at least a month within the previous year. Drug dependency includes all illicit drugs.
- Eating disorder (eating dis) = The estimated share of people with eating disorders (only includes anorexia nervosa and bulimia nervosa) in the past year, whether or not they were diagnosed, based on representative surveys, medical data and statistical modeling.
- GDP pro capita (GDP per capita) = GDP per capita is gross domestic product divided by midyear population. GDP is the sum of gross value added by all resident producers in the economy plus any product taxes and minus any subsidies not included in the value of the products. It is calculated without making deductions for depreciation of fabricated assets or for depletion and degradation of natural resources. Data are in current U.S. dollars.
- Current health expenditure (health exp) = Level of current health expenditure 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.

- Income (income) = The variable 'income' is a categorical index derived from the Global Health Expenditure Database, which classifies each country based on its gross national income per capita. Countries are divided into three income categories: 'Low', 'Middle' and 'High'. This classification provides an indication of the general level of economic development of the countries included in the dataset.

- Unemployment (unemployment) = Unemployment refers to the share of the labor force that is without work but available for and seeking employment.

- Suicide rate (suicide rate) = Annual number of suicides per 100,000 people. Suicide deaths are underreported in many countries due to social stigma and cultural or legal concerns.

- Life expectancy (life exp) = male and female life expectancy at birth.

- Urban population (urban) = Percentage of a country's total population living in urban areas.

- Internet access (internet) = Internet users are individuals who have used the Internet (from any location) in the last 3 months as a percentage of the population.

- Alcohol disorder (alcohol) = Share of the population of a country with an alcohol use disorder

- Average years of schooling (education) = Average years of education completed by the population of the country

- obesity among adults (obesity) = Percentage of adults aged 18+ with a body mass index of 30 kg/m2 or higher.

- Phones (phones)= Number of telephones owned in a certain country per 1000 inhabitants.

- Literacy (literacy) = Literacy rate in a country.

- Birthrate (birthrate) = Birth rate of a state, measured as the average number of births per year per 1000 people.

The final dataset that was used to perform the analysis comprises 158 observations, which are the countries around the world, with the dependent variable concerning the percentage of depressed people and nineteen independent variables.

Country	Depressive_dis	Anxiety_dis	Bipolar_dis	Drug_dis
Afghanistan	4.945168	4.851035	0.6996446	0.4696580
Angola	5.744194	3.934095	0.5538995	0.2951286
United Arab Emirates	3.578491	4.243272	0.7520120	0.9241267
....				
Samoa	2.802756	4.027854	0.2748158	1.0489270
Zambia	4.219490	3.969207	0.5744539	0.3284986
Zimbabwe	3.395476	3.137017	0.5385804	0.5612305

Table 1: View of the dataset

Analysing the dependent variable alone, it emerges that the countries with a higher percentage of people suffering from depression are predominantly located in Africa, with Uganda, Angola, Equatorial Guinea, Gabon, Congo and Lesotho topping this dismal list. In contrast, countries with lower values of the depression variable include Brunei (a country next to Malaysia), Singapore, South Korea, Japan, Myanmar (formerly Burma in South East Asia), Peru and Colombia. A geographical map showing the distribution of the population affected by depression by state is shown below.

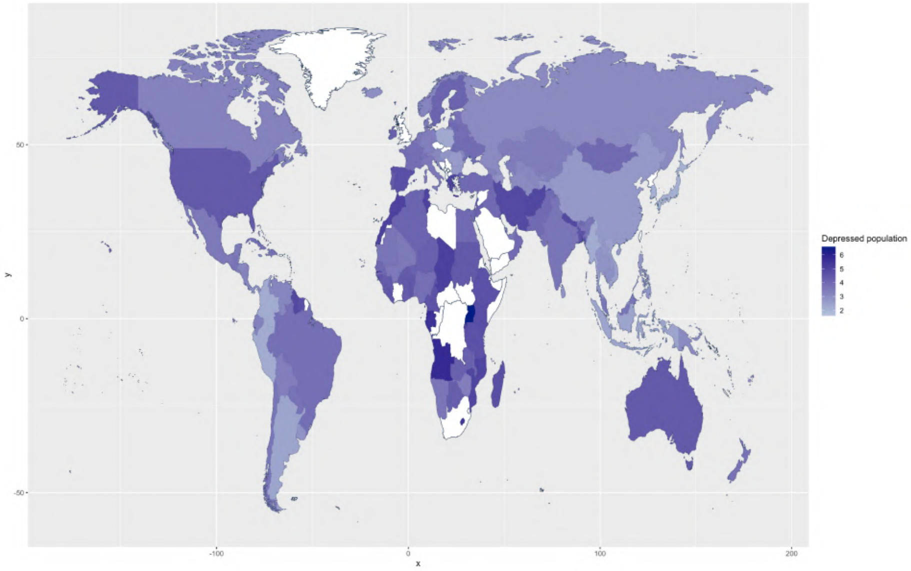


Figure 1: Worldwide distribution of depression

2 Exploratory data analysis (EDA)

The first approach that is undertaken for the analysis of statistical learning is exploratory data analysis (EDA) in which the key characteristics of individual variables are summarised using graphical representations. The primary purpose of this step is to help the reader observe the data in order to highlight its main connotations.

To examine the distribution of data, a first and widely used tool is the boxplot. This graph provides a visual representation of the distribution and centrality of the data for a continuous variable. In addition, it makes it possible to identify any anomalous values, known as outliers, which could adversely affect the final analysis. In the following, the boxplots for each continuous variable are presented, consequently excluding the variable 'Income'.

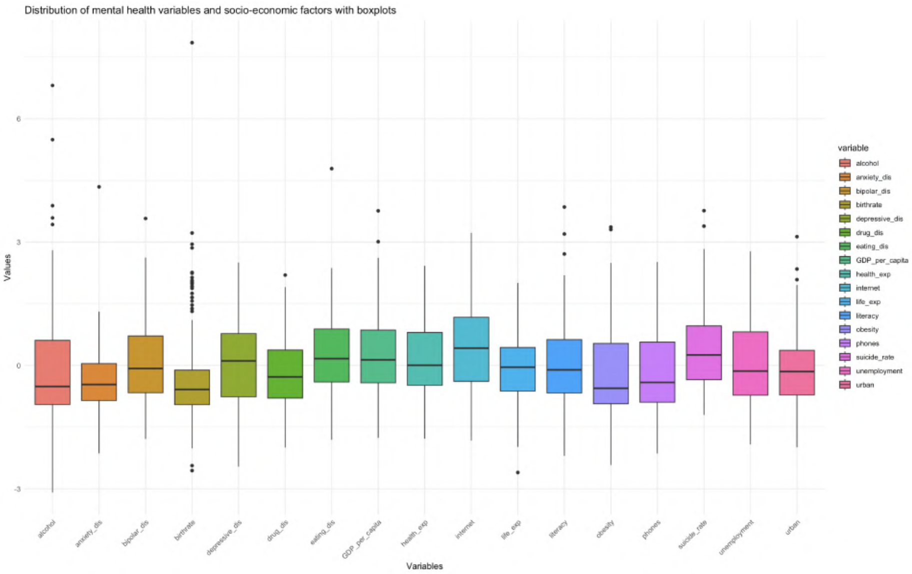


Figure 2: Distribution of mental health variables and socio-economic factors with boxplots

This graph illustrates the distribution of each numerical variable in the dataset using boxplots. This statistical tool also makes it possible to detect the presence of outliers. Starting from the left, with the variable 'alcohol' we note that there are two relevant outliers, corresponding to Mongolia and El Salvador. The problem of alcohol abuse in Mongolia is very serious and threatens to block its economic and social progress[1]. Regarding the level of anxiety in people, Portugal stands out significantly. For problems related to bipolarity we

[1] Peoples Gazette Nigeria. "Mongolia launches movement against alcoholism". 2022. https://gazettengr.com/mongolia-launches-movement-against-alcoholism/

have New Zealand as an outlier. Then, with regard to the level of births per country, there is a notable outlier that is associated with Niger[2]. On the other hand, as far as food problems are concerned, Australia is on the podium for this particular ranking. Then as far as GDP per capita is concerned, Luxembourg stands out. The outlier country that distinguishes itself negatively for life expectancy is Nigeria[3]. The two small islands of the archipelago in the South Pacific Ocean, Tonga and Samoa, are distinct with regard to the obesity problem[4]. Finally, another outlier value that emerges concerns the suicide rate. Lesotho, a small African state completely surrounded by South Africa, has the highest suicide rate in the world, with 87.5 cases per 100,000 inhabitants[5] (data taken in 2024). As we will see, we will remove these states from the dataset for a better and more truthful result.

As a second step in the exploratory data analysis, a histogram is used to examine the 15 countries with the highest incidence of depressive disorder. The graph is as follows.

[2] The Conversation. "Niger has the world's highest birth rate – and that may be a recipe for unrest". 2019. https://theconversation.com/niger-has-the-worlds-highest-birth-rate-and-that-may-be-a-recipe-for-unrest-108654

[3] Business Insider. "10 African countries with the lowest life expectancy according to the World Bank". 2024. https://africa.businessinsider.com/local/lifestyle/10-african-countries-with-the-least-life-expectancy-according-to-the-world-bank/h37db7r

[4] Daily Mail. "Obesity now greater risk to global health than hunger for first time - with 1 BILLION too fat worldwide". 2024. https://africa.businessinsider.com/local/lifestyle/10-african-countries-with-the-least-life-expectancy-according-to-the-world-bank/h37db7r

[5] The Telegraph. "'What kind of man cries?': The country with the highest suicide rate in the world". 2024. https://www.telegraph.co.uk/global-health/climate-and-people/lesotho-worlds-highest-suicide-rates-mens-mental-health-africa/

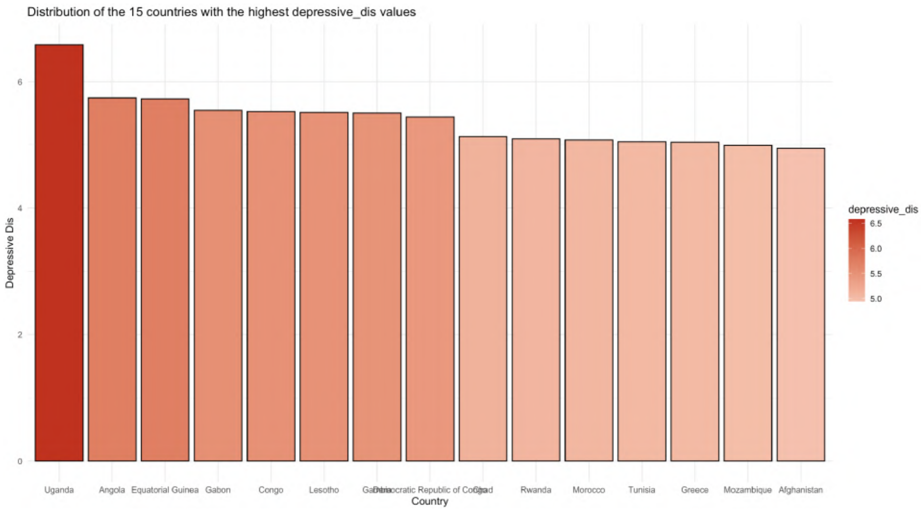


Figure 3: Distribution of the 15 countries with the highest depressive dis values

From the histogram it is possible to find in descending order, the countries with the highest percentage of people suffering from depression, Uganda, Angola, Equatorial Guinea, Gabon, Congo, Lesotho, Gambia, Democratic Republic of Congo, Chad, Rwanda, Morocco, Tunisia, Greece, Mozambique and Afghanistan. It is evident that most of these countries are in Africa, with the exception of Greece in Europe and Afghanistan in the Middle East.

In contrast, we now proceed by looking at the 15 countries in the world with the lowest percentage of population affected by depression. The histogram is shown immediately below.

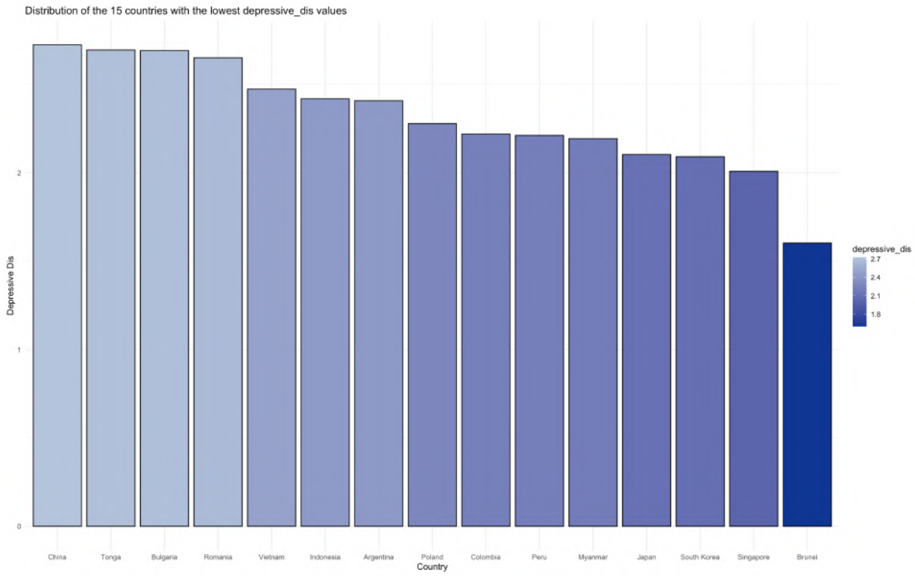
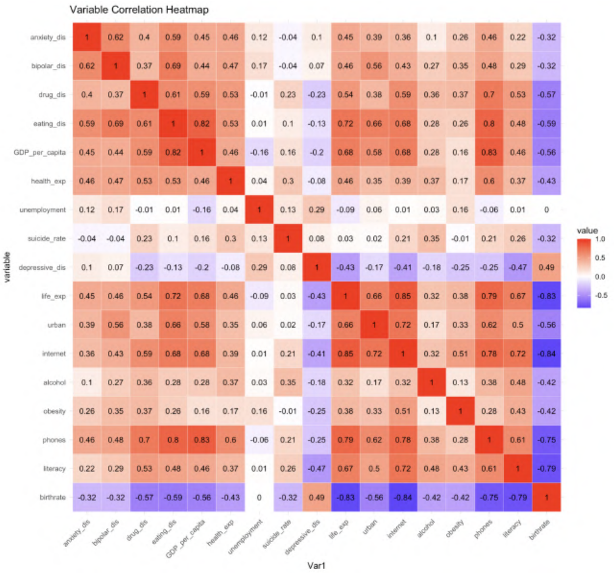


Figure 4: Distribution of the 15 countries with the lowest depressive dis values

In ascending order, the graph highlights Brunei (a small nation located near Malaysia as the country with the lowest incidence of the depression problem, followed by Singapore, South Korea, Japan, Myanmar, Peru, Colombia, Poland, Argentina, Indonesia, Vietnam, Romania, Bulgaria, Tonga and China. This list reflects a wider geographical distribution than previously observed. We note the presence of several Asian countries and the Indonesian archipelago, as well as nations from South America and Europe. As one might expect from the trend noted earlier, no African countries are included in this list.

The last step in this initial phase of analysis involves a more in-depth investigation of the correlation between the variables considered. This step is crucial as it allows us to understand whether there is some sort of linkage between pairs of factors considered and whether certain variables have a greater influence on the dependent variable, 'depressive dis'. By examining the correlations between the different variables, we are able to identify any patterns or trends that may contribute to our understanding of the underlying mechanisms of depression. Shown below is the heatmap with correlation values. The correlation index is called Pearson index, which expresses a possible relationship between two variables and it is included between the values -1 and 1.



	anxiety_dis	bipolar_dis	drug_dis	eating_dis	GDP_per_capita	health_exp	unemployment	suicide_rate	depressive_dis	life_exp	urban	internet	alcohol	obesity	phones	literacy	birthrate
anxiety_dis	1	0.62	0.4	0.59	0.45	0.46	0.12	-0.04	0.1	0.45	0.39	0.36	0.1	0.26	0.46	0.22	-0.32
bipolar_dis	0.62	1	0.37	0.69	0.44	0.47	0.17	-0.04	0.07	0.46	0.56	0.43	0.27	0.35	0.48	0.29	-0.32
drug_dis	0.4	0.37	1	0.61	0.59	0.53	-0.01	0.23	-0.23	0.54	0.38	0.59	0.36	0.37	0.7	0.53	-0.57
eating_dis	0.59	0.69	0.61	1	0.82	0.53	0.01	0.1	-0.13	0.72	0.66	0.68	0.28	0.26	0.8	0.48	-0.59
GDP_per_capita	0.45	0.44	0.59	0.82	1	0.46	-0.16	0.16	-0.2	0.68	0.58	0.68	0.28	0.16	0.83	0.46	-0.56
health_exp	0.46	0.47	0.53	0.53	0.46	1	0.04	0.3	-0.08	0.46	0.35	0.39	0.37	0.17	0.6	0.37	-0.43
unemployment	0.12	0.17	-0.01	0.01	-0.16	0.04	1	0.13	0.29	-0.09	0.06	0.01	0.03	0.16	-0.06	0.01	0
suicide_rate	-0.04	-0.04	0.23	0.1	0.16	0.3	0.13	1	0.08	0.03	0.02	0.21	0.35	-0.01	0.21	0.26	-0.32
depressive_dis	0.1	0.07	-0.23	-0.13	-0.2	-0.08	0.29	0.08	1	-0.43	-0.17	-0.41	-0.18	-0.25	-0.25	-0.47	0.49
life_exp	0.45	0.46	0.54	0.72	0.68	0.46	-0.09	0.03	-0.43	1	0.66	0.85	0.32	0.38	0.79	0.67	-0.83
urban	0.39	0.56	0.38	0.66	0.58	0.35	0.06	0.02	-0.17	0.66	1	0.72	0.17	0.33	0.62	0.5	-0.56
internet	0.36	0.43	0.59	0.68	0.68	0.39	0.01	0.21	-0.41	0.85	0.72	1	0.32	0.51	0.78	0.72	-0.84
alcohol	0.1	0.27	0.36	0.28	0.28	0.37	0.03	0.35	-0.18	0.32	0.17	0.32	1	0.13	0.38	0.48	-0.42
obesity	0.26	0.35	0.37	0.26	0.16	0.17	0.16	-0.01	-0.25	0.38	0.33	0.51	0.13	1	0.28	0.43	-0.42
phones	0.46	0.48	0.7	0.8	0.83	0.6	-0.06	0.21	-0.25	0.79	0.62	0.78	0.38	0.28	1	0.61	-0.75
literacy	0.22	0.29	0.53	0.48	0.46	0.37	0.01	0.26	-0.47	0.67	0.5	0.72	0.48	0.43	0.61	1	-0.79
birthrate	-0.32	-0.32	-0.57	-0.59	-0.56	-0.43	0	-0.32	0.49	-0.83	-0.56	-0.84	-0.42	-0.42	-0.75	-0.79	1

Figure 5: Heatmap with correlations between variables

The correlation matrix reveals that the variable 'depressive dis' (depression index) shows a significant correlation with several other variables in the dataset. In particular, there is a moderate positive correlation with the variable 'birthrate', which implies that countries with a higher birthrate tend to show higher levels of depression. On the other hand, 'depressive dis' shows significant negative correlations with variables such as 'life exp' (life expectancy), 'internet' (internet access), 'literacy' (literacy rate) and 'phones' (telephone penetration), suggesting that countries with higher life expectancy, internet access, literacy rates and telephone penetration tend to show lower levels of depression.

However, it is important to note that the correlation between some variables could indicate the presence of multicollinearity, which could affect the stability and interpretation of the linear regression model. For example, the strong correlation between "internet" and "phones" could suggest that these variables provide similar information, which could cause multicollinearity problems in the model.

3 Unsupervised Learning

In the context of data analysis and for the purpose of the investigation that is being carried out, the study moves on to the part on unsupervised learning. This important statistical learning procedure is crucial for exploring and obtaining useful information regarding the data set without already defined labels. Through this approach, it is possible to discover hidden trends and patterns in the data without depending on predefined output information. In order to obtain results within this investigation, we will focus on the application of this learning modality to better analyse data related to mental health and some socio-economic aspects.

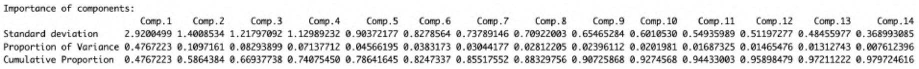
The two main techniques of principal component analysis (PCA) and clustering will be used. PCA will make it possible to reduce the dimensionality of the dataset and identify the most significant patterns. Clustering, on the other hand, is in turn subdivided into hierarchical clustering and k-means to obtain agglomerates of distinct states but with similar observations.

In the course of the following chapter, these two unsupervised learning techniques will be applied in order to delve into the data we possess in greater detail in order to develop an even more comprehensive view of our investigation.

3.1 PCA

Principal Component Analysis (PCA) is an essential methodology for analysing multivariate data. It allows us to obtain a low-dimensional representation of our dataset, reducing the complexity of the data while retaining most of the relevant information. This technique identifies an optimal linear combination of the original variables that maximises the variance of the data and are uncorrelated with each other. In this way, principal components are generated that efficiently represent the variances present in the original dataset. PCA allows the identification of the main directions along which the data vary the most and effectively represents the underlying structure of the data in a space of reduced size.

The first step is to normalise the data and remove the categorical variable 'Income' from the study in order not to misinterpret the numerical values. The data were normalised because the different variables have different scales. This avoids the fact that some variables may have a more significant dominance than others. Then the correlation matrix between the different variables is calculated (as was done in the previous chapter). The covariance matrix could also be used, but we will use the former instead, as correlation is the most commonly used method. As mentioned earlier, the level of correlation is denoted by the Pearson index and can vary from -1 (inverse correlation) to 1 (positive correlation). After having performed these initial steps to prepare the process, we have all the elements at our disposal to conduct the principal component analysis. The first thing to understand is the number of components to optimally synthesise the available dataset and explain as much variance as possible from the starting variables. Several approaches can be used to figure out the optimal number of components to use. The first approach described is that the number of components is chosen from the point of view of explaining the percentage of total variance. The output on the importance of the components is as follows.



```
Importance of components:
                          Comp.1    Comp.2     Comp.3     Comp.4     Comp.5    Comp.6     Comp.7     Comp.8     Comp.9   Comp.10    Comp.11    Comp.12    Comp.13     Comp.14
Standard deviation     2.9200499 1.4008534 1.21797092 1.12989232 0.90372177 0.8278564 0.73789146 0.70922003 0.65465284 0.6010530 0.54935989 0.51197277 0.48455977 0.368993085
Proportion of Variance 0.4767223 0.1097161 0.08293899 0.07137712 0.04566195 0.0383173 0.03044177 0.02812205 0.02396112 0.0201981 0.01687325 0.01465476 0.01312743 0.007612396
Cumulative Proportion  0.4767223 0.5864384 0.66937738 0.74075450 0.78641645 0.8247337 0.85517552 0.88329756 0.90725868 0.9274568 0.94433003 0.95898479 0.97211222 0.979724616
```

Figure 6: Importance of Components

The analysis obtained shows that 18 components were generated, corresponding to the number of variables present in the dataset. Each component, as indicated by the "Proportion of Variance" section, illustrates the percentage of total variance present in the observations. The formula 0. 95 raised to p, all times 100, can be used to choose the correct number of components. This formula means that the extracted components take into account on average at least 95% of the variance of each of the p starting variables. In our case p = 18 and the variance to be explained is equal to 39.72%. Thus, we have to choose the number of components that explain 39.72% of the total variance. For example, the first component explains 47.67% of the total variance, indicating that approximately 48% of the data of the 18 variables can be represented by this component alone. The second component, on the other hand, explains 10.97% of the total variance. The "Cumulative Proportion" section shows that the first two components explain 58.64% of the total variance of the observations. Subsequently, each component will contribute decreasingly to the total variance, until a cumulative variance of 1 is reached, which is 100% for the last component.

According to the written formula, the first two components largely explain the part of the cumulative variance calculated above. Therefore, the optimal number of components for this analysis is 2.

Furthermore, in order to graphically demonstrate what has just been stated, a scree-plot is shown below in which the explanation of the total variance is shown on the vertical axis and the number of different components on the horizontal axis.

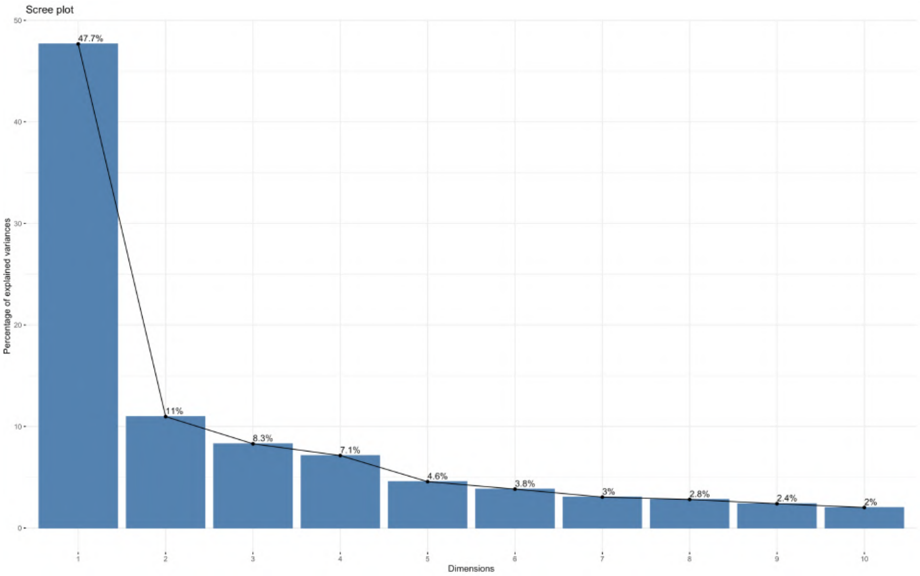


Figure 7: Scree plot with the percentage of variance explained

To continue the investigation, we proceed with the use of the scree-plot in which the eigenvalues are shown on the vertical axis. This graph shows the eigenvalues for each individual component. In order to understand the correct number of components for this graph, the "Elbow method" is used, which means that the number of components to be extracted is the one that coincides with the change in slope of the spline. As the red straight line also shows, the change in slope corresponds to the second component.

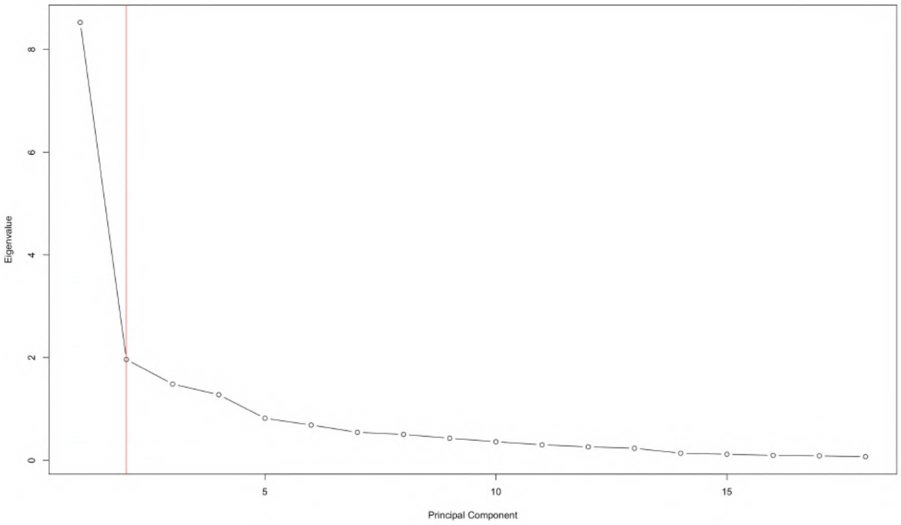


Figure 8: Scree plot with eigenvalues

Finally, having made these arguments explicit and through a mixture of the two methods used, it can be stated that the optimal number of components is 2.

A graph called a biplot will be used to graphically represent the principal component analysis. It is shown below.

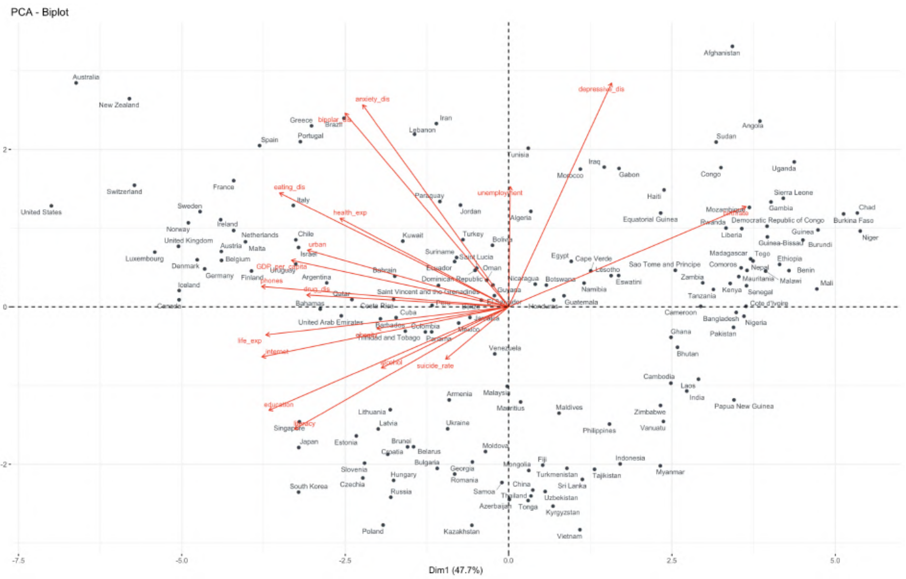


Figure 9: Biplot for PCA

From the graph above, some interpretations can be made, for example, regarding the correlation of the variables with the main components. The second component can be identified as a synthetic indicator of social problems as it is positively correlated with the variables 'depressive dis', 'unemployment', 'anxiety dis', 'bipolar dis' and inversely with 'suicide rate' or 'alcohol'. In contrast, the first component is both an indicator of social and economic aspects of a nation and of the individual as it is inversely correlated with the remaining variables.

Looking at the graph, we can make interpretations on groups of variables. For example, if we observe that the variables 'GDP per capita', 'urban', 'internet', 'health expenditure', 'education' and 'phones' have similar weights on the first principal component, we might suggest that these variables are correlated with each other in terms of economic development and technological infrastructure. Similarly, if other variables such as "anxiety dis", "bipolar dis", "eating dis", "alcohol" and "obesity" show similar weights on the second principal component, we might hypothesise that they are correlated in terms of mental health or psychological well-being.

To provide an example for each of the groups, for the first group, we examine Finland. Finland's position among these variables suggests a high level of economic development, good technological infrastructure and high living standards, which could indicate a high level of general well-being in the country. Similar conclusions can be reached for other European countries such as Belgium, Austria, Germany, Denmark, Iceland, Sweden and others.

Next, looking at the length of the arrows associated with each variable in the main components, it can be seen that most of the variables are satisfactorily explained by the structure of the data. However, it is evident that the variables 'suicide rate', 'unemployment', 'alcohol' and 'obesity' are less effectively explained, as its arrows appear shorter, suggesting a lower association with the other variables within the main components.

To conclude, analysing the graph, it can be stated that there are some values that deviate significantly from the others and, therefore, could alter the results in future analysis procedures. Two in particular emerge, the United States and Afghanistan (In addition to the aforementioned Australia and New Zealand). These two observations in the supervised part of the analysis will be eliminated in order to obtain undistorted and, therefore, better results.

3.2 Clustering

After having proceeded with the principal component analysis, we move on to the implementation of clustering. This technique aims at selecting and assembling groups of observations, specifically called clusters. These clusters group observations through similarity between them, calculated in terms of distance. We will focus on two main clustering techniques: k-means clustering and hierarchical clustering.

3.2.1 K-Means Clustering

K-Means clustering is an unsupervised learning procedure in which groups of observations are created on the basis of their distances. In our case, states are randomly assigned to the initially defined K clusters. The distance between the individual element and the centroid (midpoint of the clusters) is calculated. The centroids, in an iterative process continue to recalculate until convergence occurs, and the observations no longer change clusters.

The first step is to choose the value of K, that is, what is the optimal number of clusters to have. In this analysis, we will employ three main techniques for this function, the Within Sum of Squares Plot, the Silhouette and the NbClust.

The Within Sum of Squares (WSS) plot to determine the value of K, shows the trend of the total square error as K changes. In order to work out the correct number of clusters, one must reason by means of the elbow method in which one chooses the point at which the function decreases in slope. In our case, as can be seen from the graph below, the optimal number K is 3.

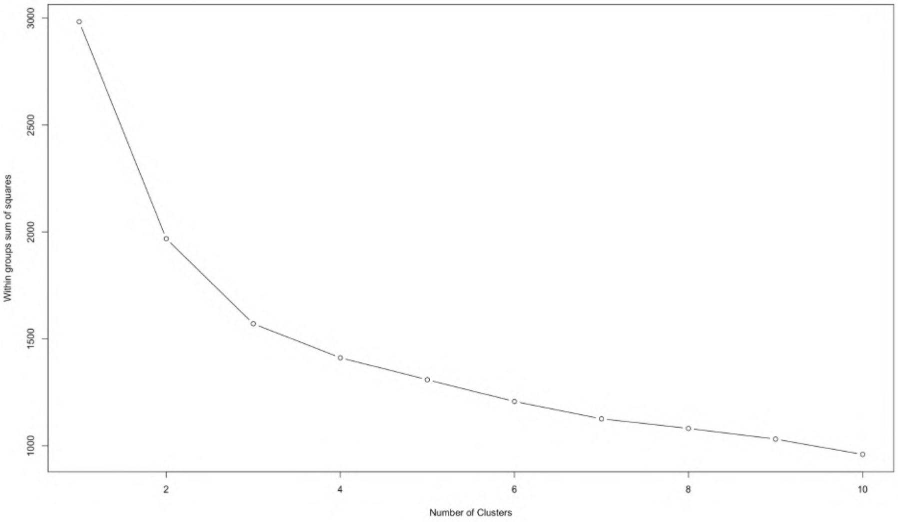


Figure 10: Within sum of squares plot

In order to confirm what has just been demonstrated, the Silhouette method is used. The silhouette value shows how similar the individual element is to the cluster it belongs to in relation to the others. The value ranges from -1 to 1. If the average score tends towards 1, it means that the cluster possesses very similar states. As the value decreases, the compactness of the cluster also decreases. Furthermore, negative values mean that the observations considered should have been included in another, more similar cluster. The two graphs most relevant to the study confirm that it is optimal to choose K equal to 3. In fact, the average value of the silhouette is higher (0.25 on the left and 0.24 on the right).

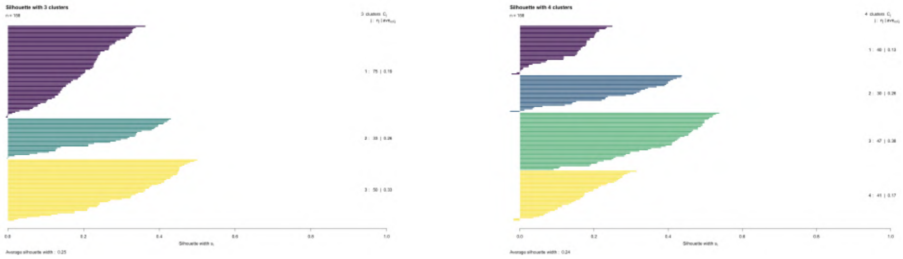
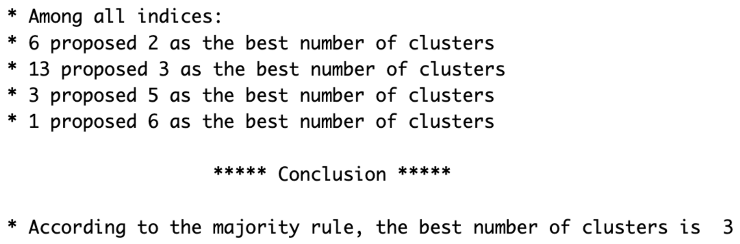


Figure 11: Comparison of silhouettes with 3 and 4 clusters

Finally, the last technique is to process the NbClust function on the R programme to determine the perfect number of clusters. The final output states, as can be seen from the figure, that the automatic function suggests the use of 3 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
```

Figure 12: NbClust function

After deciding on the perfect number of clusters to be used, we proceed to the actual implementation of the clustering process, partitioning the observations into 3 main groups according to the similarity criterion. The partitioning graph is as follows.

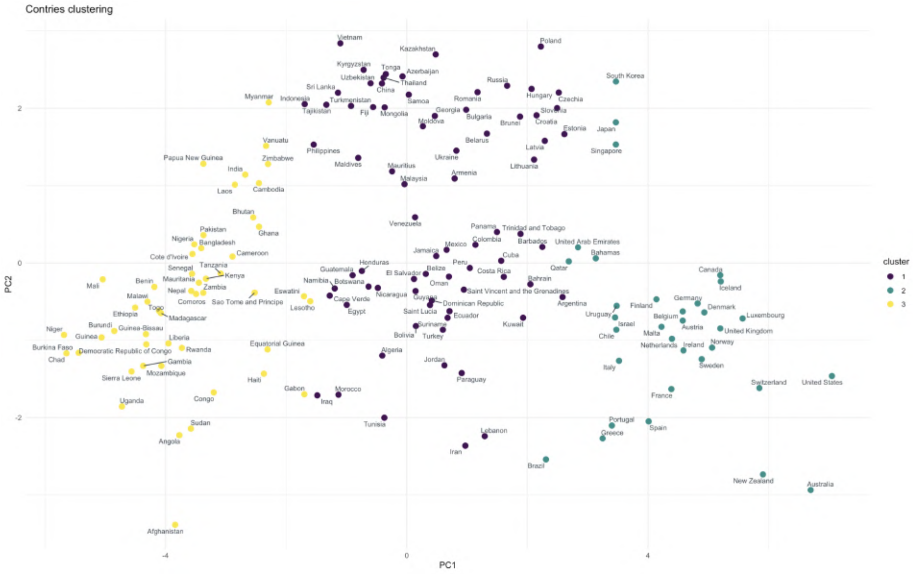


Figure 13: K-Means clustering

For an easier and more immediate visualisation, a geographical map is also presented in which the states are divided by cluster colour.

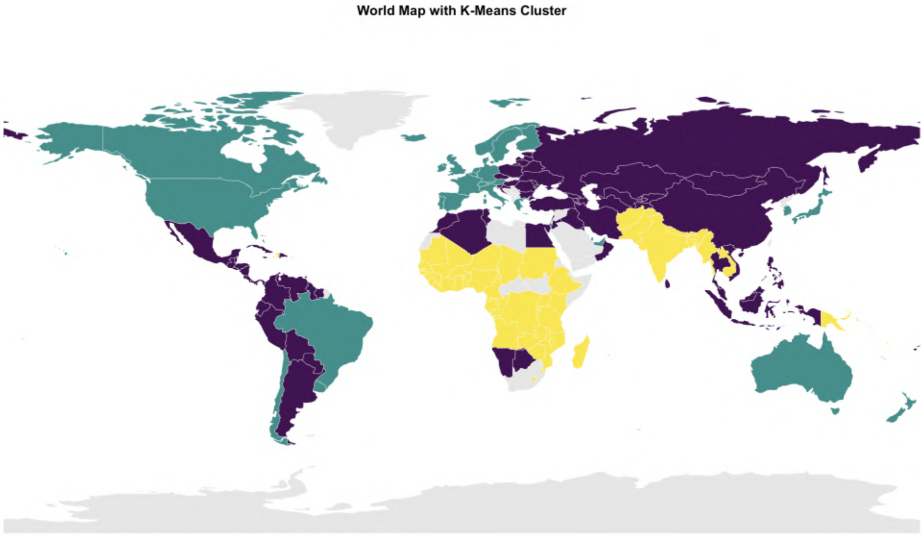
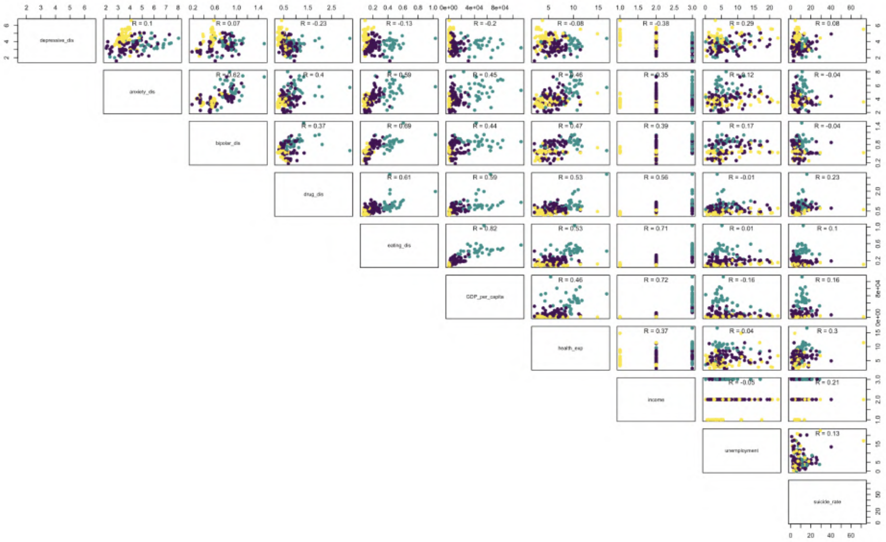
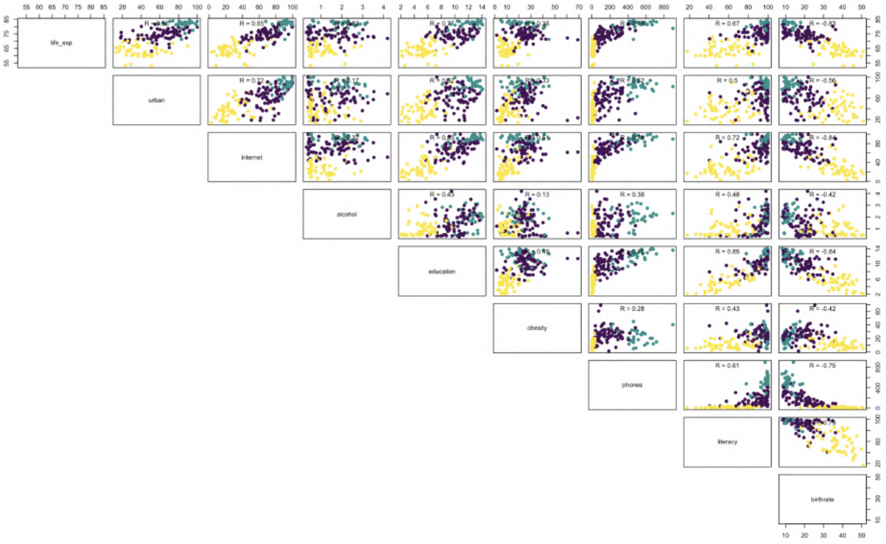


Figure 14: World map with k-means clustering

Analysing the three clusters, number 2 coloured in green stands out first of all. This cluster includes a group of countries geolocated mainly in Western Europe, North America, Australia, New Zealand, some South American countries, Japan, South Korea and Singapore. Also included are two eastern countries that are distinguishing themselves through their investments for the occidental world such as Qatar and the United Arab Emirates. The first cluster, represented in purple, includes countries mainly located in Asia, Eastern Europe, Central and South America, North Africa, and the Indonesian archipelago. The third cluster, identified by the colour yellow, emerges with a less dispersed geographical distribution. It involves all countries in Central Africa, South Asia (including India), Papua New Guinea in Oceania and Haiti in the Caribbean.

Using the following scatterplots, some interpretations can be made.





The countries of the second clusters are characterised by having a higher GDP per capita, higher investment in healthcare, higher education, urbanisation and life expectancy, and lower unemployment. Despite this, these countries are also characterised by having a population affected by certain mental health problems. In fact, they have a higher percentage of people with anxiety problems, bipolarism, eating disorders, alcoholism and drug use. They also have the highest suicide rate on average and are the countries with the lowest birth rate. Therefore, there is a contrast in these countries in that they are the richest countries economically, but have significant social problems. Not surprisingly, several studies have tried to observe this phenomenon, and many researchers confirm this[6]. The countries of the first cluster show average values in all variables except obesity. This suggests that they are in an intermediate position between the economically richer and poorer ones. In fact, economically, they are generally classified as middle-income countries. Furthermore, with regard to obesity, some research indicates that in these middle-income countries the problem of overweight is increasing more rapidly than in richer countries[7]. The countries of the third cluster have the highest level of 'depressive dis', the dependent variable under investigation. However, it is remarkable to note that other mental health indicators, such as anxiety, bipolarism, drug and alcohol use, as well as eating disorders and suicide rates, register lower values. These countries also have a lower GDP per capita and invest significantly less in key areas such as education, health, urbanisation, internet access and technology than the others. It is relevant to note that they are also the countries with the highest birth rate, but lower life expectancy. To better graphically show some significant variables, six density graphs are shown.

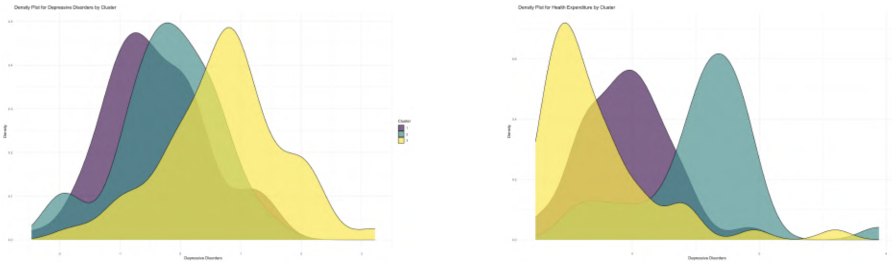


Figure 15: Density plot for depressive disorders and GDP per capita by clusters

[6] Los Angeles Times. "Depression higher in wealthy nations, research suggests". 2011. https://www.latimes.com/health/la-xpm-2011-jul-26-la-heb-depression-wealthy-countries-20110726-story.html

[7] Le Monde. "Obesity catches up with low- and middle-income countries". 2023. https://www.lemonde.fr/en/health/article/2023/07/25/obesity-catches-up-with-low-and-middle-income-countries_6067132_14.html

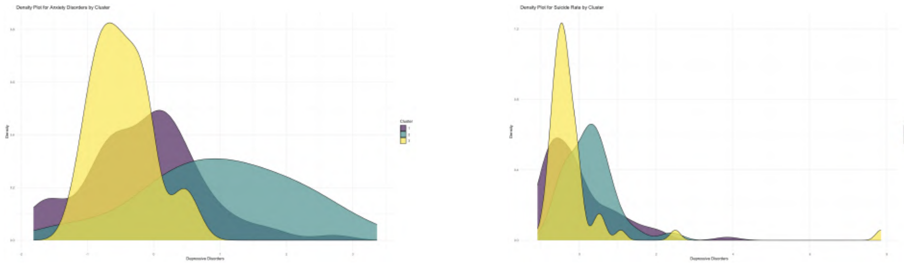


Figure 16: Density plot for anxiety disorders and suicide rate by clusters

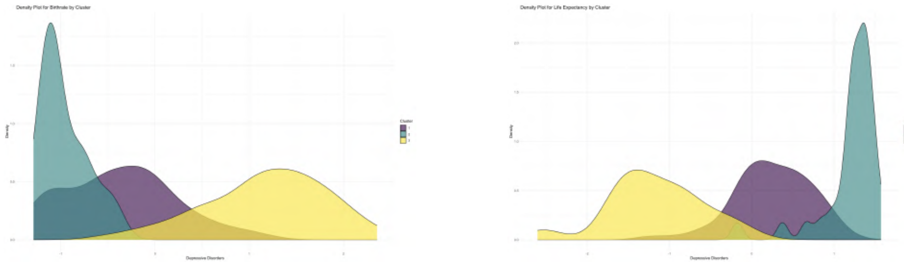


Figure 17: Density plot for birthrate and life expectancy by clusters

We have concluded the phase on the k-means clustering algorithm. Through this process, we came up with intriguing results. For example, we observed that economically and technologically advanced countries exhibit more mental health problems among their citizens, while less developed countries show a clear disparity between birth rate and life expectancy.

3.2.2 Hierarchical Clustering

The second step with regard to clustering concerns another approach, the hierarchical clustering. This produces groups of observations by exploiting an iterative hierarchical process. At first, each observation is a separate cluster. Then, the clustering of the most similar pairs proceeds iteratively until all observations are all under the same cluster. The easiest way to visualise the result is to create a tree graph called a dendrogram. The dendrogram reproduces the hierarchical distribution of each observation.

First of all, the dissimilarity matrix must be calculated. Then it must be specified with which linkage method the individual observations should be associated. There are several ways to obtain the linkage between the states. In this study, three different ways will be analysed, the average linkage, the complete linkage and the Ward linkage. The method of linkage that produces a graphically better output will be chosen. In addition, 5 clusters per partition will be chosen automatically by default.

By averaging the linkage, the average of the distinct distances between the observations is calculated and the dendrogram obtained is reported below.

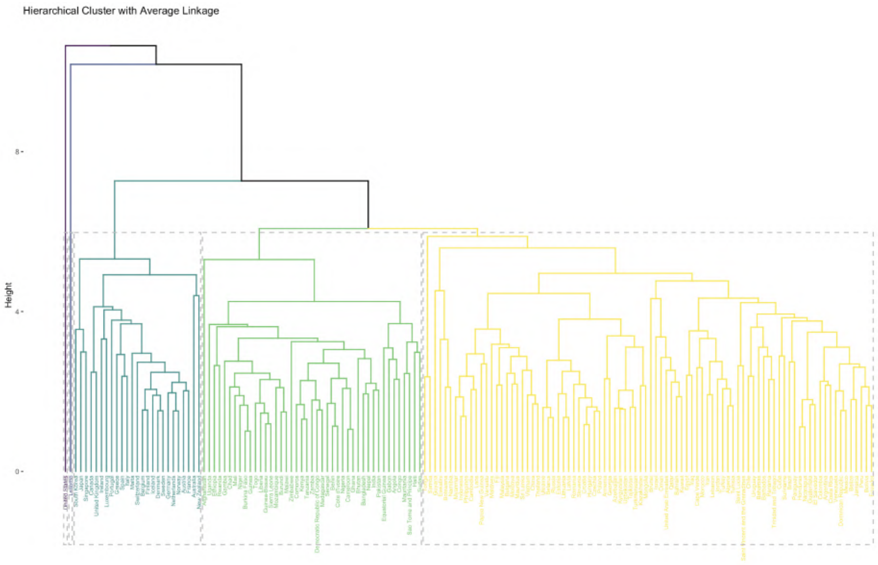


Figure 18: Hierarchical cluster with average linkage

It can be seen from the graph that three clusters of a good size were created. Unfortunately, the first two clusters from the left have only one observation each, the United States and Lesotho, respectively. Therefore, it can be asserted that these two countries have a very large distance to the other states and, consequently, are not similar. For this reason, as stated in the previous paragraphs, both observations will be eliminated in order to obtain an unbiased and more truthful result.

The next method that will be used to observe the relationships between observations will be through complete linkage. In this case, the similarity between two groups is defined as the largest value of distance between the observations of the two groups. The dendrogram is below.

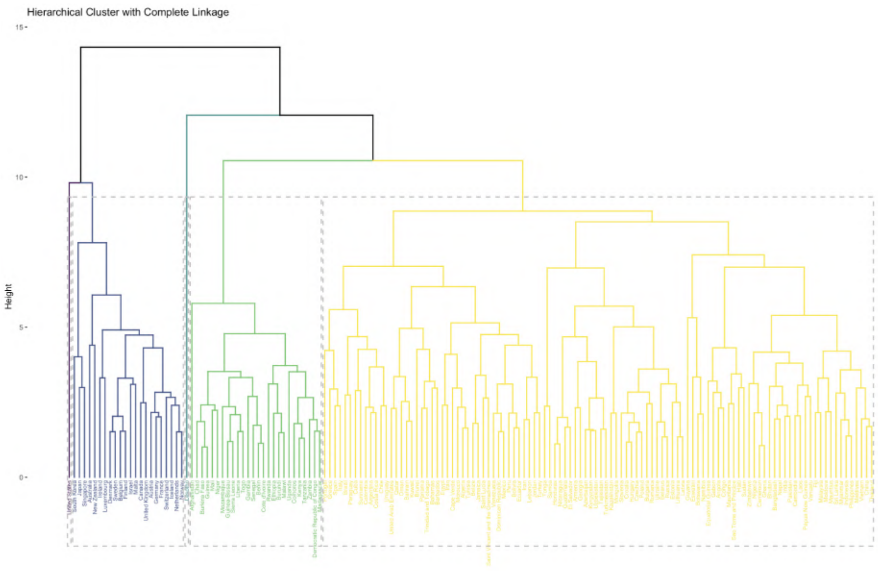


Figure 19: Hierarchical cluster with complete linkage

The graph shows a predominance in the size of the last cluster from the left (in yellow). Furthermore, as in the previous case, there are two clusters corresponding to a single observation, United States and Lesotho. To try to avoid this problem and obtain clusters of a homogeneous size we try to use another type of linkage.

The last type of linkage that will be used to calculate similarities between nations is the Ward's linkage. The latter combines groups that obtain the lowest growth of a given measure of heterogeneity. Thus, this linkage can be said to minimise the total within-group variance. The dendrogram, as we are going to see, will give a more organic presentation of the distribution of states.

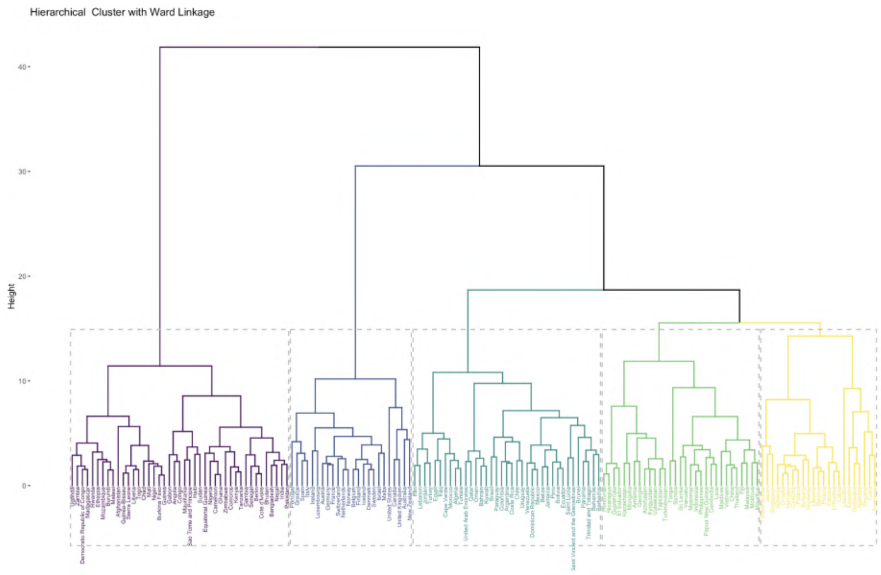


Figure 20: Hierarchical cluster with Ward linkage

As can be seen at first glance, the distribution of observations is more amalgamated. Five clusters have been created that include a more homogeneous number of states. In fact, starting from the left, the first cluster (in purple) includes 43 states. The fourth cluster (in blue) includes 24 observations. Also from the left, the second cluster (in dark green) consists of 37 observations. Then the third cluster (in light green) includes 31 observations. Finally, the fifth cluster (in yellow) has 23 observations.

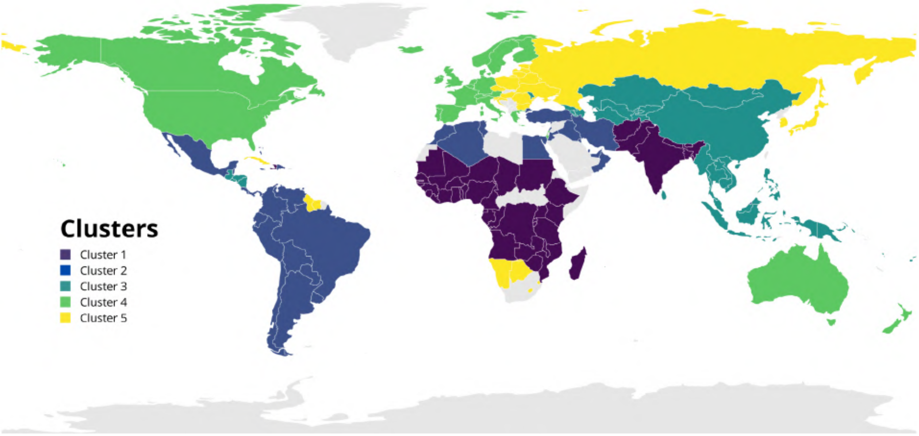
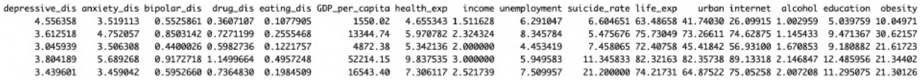


Figure 21: World map with k-means clustering

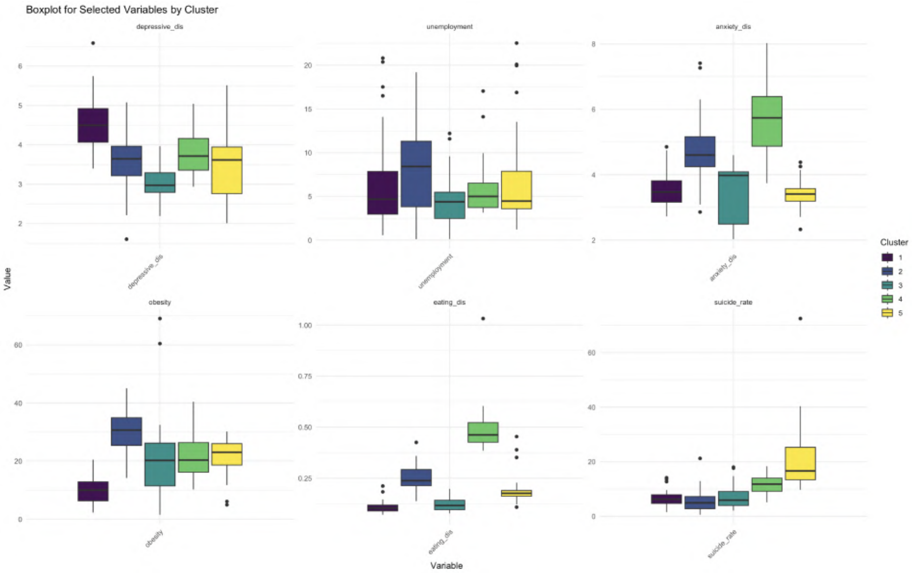
Hierarchical clustering allows a better distinction than k-means clustering. An interpretation of the results by averaging the values for each variable follows.



```
depressive_dis anxiety_dis bipolar_dis  drug_dis eating_dis GDP_per_capita health_exp   income unemployment suicide_rate life_exp    urban internet  alcohol education  obesity
      4.556358    3.519113   0.5525861 0.3607107  0.1077905        1550.02   4.655343 1.511628     6.291047     6.604651 63.48658 41.74030 26.09915 1.002959  5.039759 10.04971
      3.612518    4.752057   0.8503142 0.7271199  0.2555468       13344.74   5.970782 2.324324     8.345784     5.475676 75.73049 73.26611 74.62875 1.145433  9.471367 30.62157
      3.045939    3.506308   0.4400026 0.5982736  0.1221757        4872.38   5.342136 2.000000     4.453419     7.458065 72.40758 45.41842 56.93100 1.670853  9.180882 21.61723
      3.804189    5.689268   0.9172718 1.1499664  0.4957248       52214.15   9.837535 3.000000     5.949583    11.345833 82.32163 82.35738 89.13318 2.146847 12.485956 21.34402
      3.439601    3.459042   0.5952660 0.7364830  0.1984509       16543.40   7.306117 2.521739     7.509957    21.200000 74.21731 64.87522 75.05258 2.007208 11.295075 21.30126
```

Figure 22: Means

The first cluster is very similar to the third cluster of k-means clustering because it includes most of the central countries of Africa and South Asia. In terms of mental health, these countries are characterised by having the highest percentage of people suffering from depression. Despite this, they have fewer problems related to nutrition, such as obesity and drug and alcohol abuse. Economically, urbanistically, technologically and educationally, these countries show significantly lower indicators than the others. This suggests that they can be considered among the most socio-economically disadvantaged within the dataset. The second cluster includes most of the countries of South America, Mexico, the countries of North Africa and the Middle East. These countries have an average income. In addition, they are distinguished by having high rates of unemployment, accompanied by a higher incidence of suicide and obesity. The third cluster, highlighted in dark green, comprises Central Asian countries, including China, the Indonesian archipelago states and some Central American countries. These middle-income countries have lower levels of depression and bipolarity than the others. They also have the highest percentage of employed population. The fourth cluster in light green consists of the countries of Western Europe, North America and Oceania. These countries stand out for a significant contrast: while they have higher levels of anxiety, drug and alcohol use disorders, bipolarism and eating disorders, they also have higher incomes and invest more in key areas such as education, urbanisation, technology and health. Moreover, despite mental health problems, they have a longer life expectancy than other countries. Moreover, life expectancy is among the highest in the world. Despite this, they are the countries with the lowest birth rates. The fifth cluster, marked by the colour yellow, includes Russia, the Eastern European countries (belonging to the old Soviet bloc) and other nations scattered across Southern Africa and Latin America. These countries are positioned at a medium-rich level of wealth. However, they are distinguished by their surprisingly high suicide rates, which are significantly higher than in the other clusters. To better understand the distribution of some relevant variables (depressive dis, unemployment, anxiety dis, obesity, eating dis and suicide rate), boxplots are shown.


Boxplot for Selected Variables by Cluster
depressive_dis
6
5
4
3
2
depressive_dis
unemployment
20
15
10
5
0
unemployment
anxiety_dis
8
6
4
2
anxiety_dis
Cluster
1
2
3
4
5
Value
obesity
60
40
20
0
obesity
eating_dis
1.00
0.75
0.50
0.25
eating_dis
suicide_rate
60
40
20
0
suicide_rate
Variable

4 Supervised Learning

The first part of the analysis has been completed. The study on the second, the supervised part, is now being carried out. Supervised learning is a statistical learning approach in which computer algorithms are used to detect relationships between input data and output variables in the available datasets. Thus, it differs from the unsupervised approach, in which there was no reference output between the data. The objective of supervised learning is to build a model capable of making predictions on new data based on the relationships learned from the training data.

This section aims to identify the most relevant variables in determining the percentage of the population affected by depression in a country. To achieve this goal, we will adopt several analytical techniques. First, we will use linear regression, implemented manually to have direct control over the model training process. Next, we will explore the stepwise regression approach, which automatically selects the most informative variables. Furthermore, we will consider ridge and lasso regression, which introduce coefficient penalties to handle multicollinearity and improve the generalisation of the model. Finally, we will examine the decision tree, an intuitive model that can reveal non-linear relationships between variables and the target.

4.1 Assumptions and Linear Regression

Before carrying out any supervised learning model, it is essential to deal with four main assumptions: removing outliers, checking the homogeneity of variables and residuals and assessing multicollinearity. We then proceed to remove the outliers that were detected in the previous chapters during the exploratory data analysis and the unsupervised part. In the first chapter, we observed a large difference in specific variables for Mongolia, El Salvador, Portugal, New Zealand, Niger, Australia, Luxembourg, Nigeria, Tonga, Samoa and Lesotho. Then, in the PCA and the hierarchical clustering process, it was observed that the United States and Afghanistan were disconnected. Therefore, these states were removed from the dataset as they may negatively affect the accuracy of the model. Removing them is the best way to obtain more reliable and meaningful results.

The second assumption we have to assess is the homogeneity of the variables. Therefore, it is necessary to examine the distribution of the variables and analyse whether they follow a normal distribution. We proceed to analyse the dependent variable 'depressive dis' as an example. A graph will be used that plots the quantiles of the variable against the theoretical quantiles of the normal distribution. If the points on the graph tend to follow the bisector, then there is a normal distribution. In addition, a density graph that overlays a dashed curve of normal density on the data distribution is also presented. This helps to better visualise if the data deviates from the normal distribution.

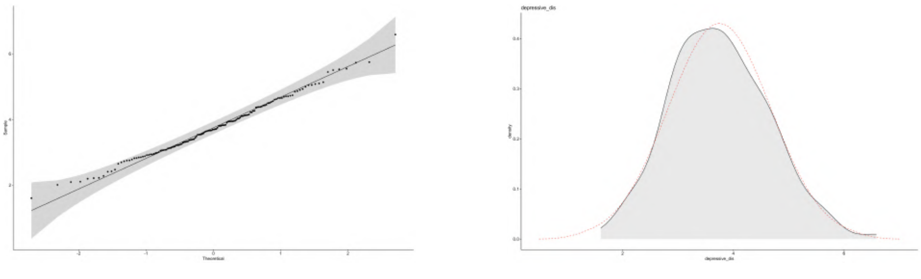


Figure 23: QQ plot and density distribution of the variable depressive dis

The distribution of data on persons suffering from depression seems to follow a trend towards a normal distribution. This observation is particularly relevant for the analysis we are about to conduct. Making the same graphs for all variables, it can be observed that a total of five variables do not follow a normal distribution, such as 'eating dis', 'GDP per capita', 'unemployment', 'suicide rate' and 'literacy'. One way to overcome this problem is to try making the variables logarithmic. In this way, the distribution can be more similar to the normal distribution, as it tends to compress the larger values and expand the smaller ones. With regard to the variable on people with food-related problems, GDP per capita, unemployment, suicide rate there is a considerable increase in the results, as can be seen from the two comparative graphs (left before and right after the logarithmic transformation).

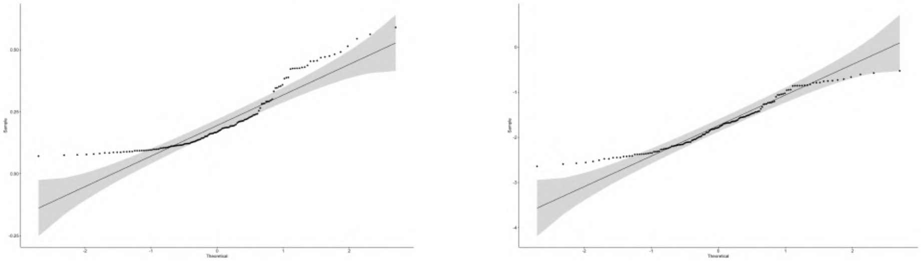


Figure 24: Before and after the logarithmic transformation for eating dis

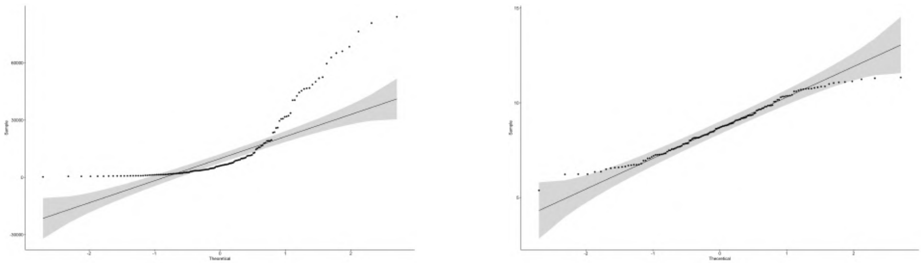


Figure 25: Before and after the logarithmic transformation for GDP per capita

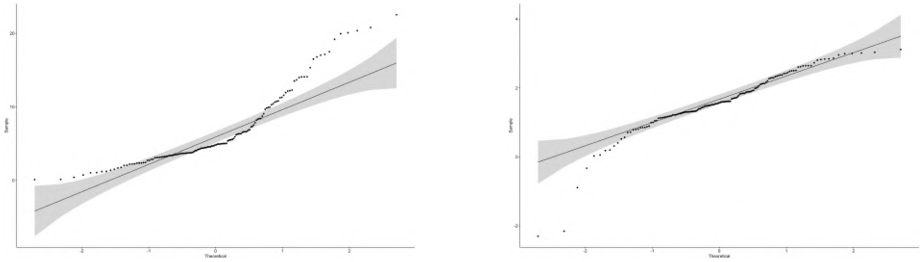


Figure 26: Before and after the logarithmic transformation for unemployment

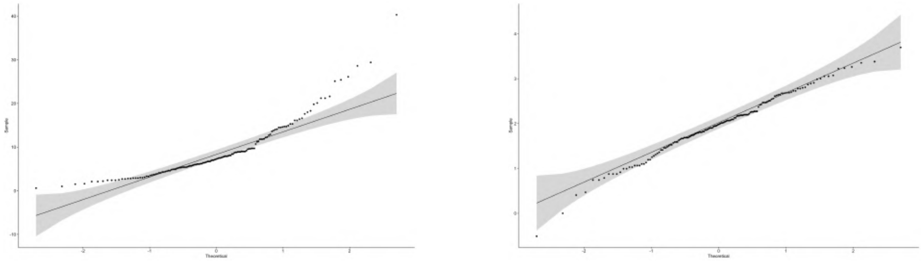


Figure 27: Before and after the logarithmic transformation for suicide rate

With regard to the variable 'literacy', no improvement in the distribution is observed (also from the graphs) even after exploring possible transformations. Therefore, for the forthcoming analyses, we choose not to consider this variable.

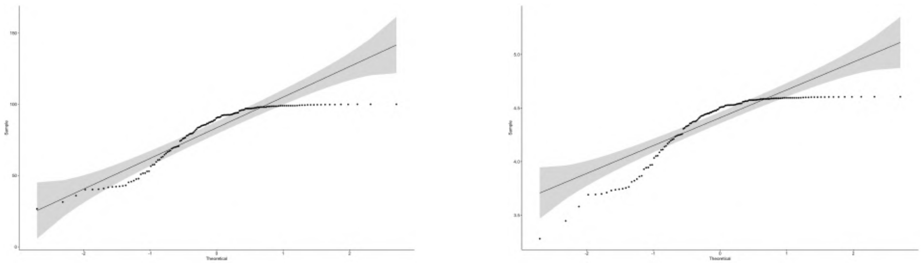


Figure 28: Before and after the logarithmic transformation for literacy

In a linear regression analysis, it's crucial to ensure that the residuals, which are the differences between observed and predicted values, follow a normal distribution with a mean of 0. This is a key assumption for the validity of the model. The histogram depicted below demonstrates that the residuals exhibit a symmetrical distribution centered around 0. This symmetry suggests that the residuals adhere to a normal distribution, validating one of the fundamental assumptions of linear regression.

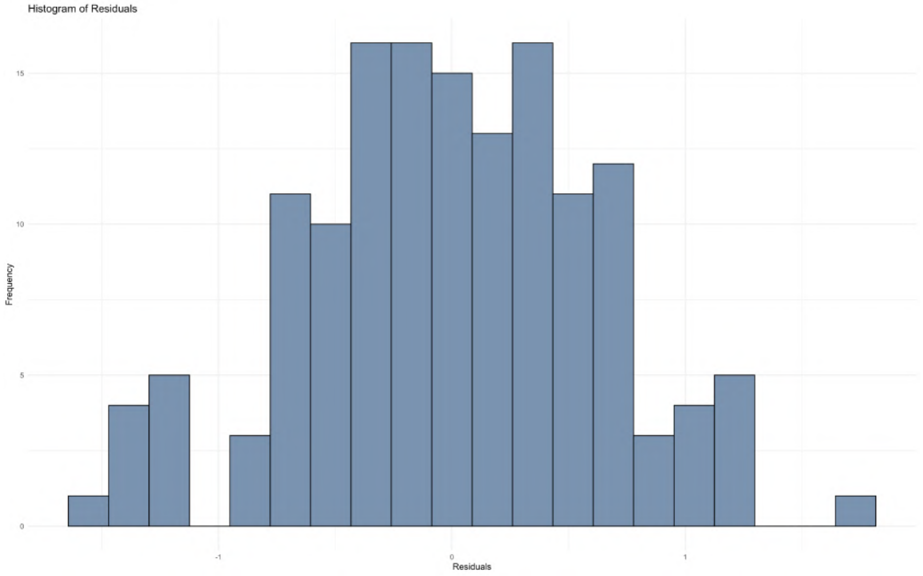
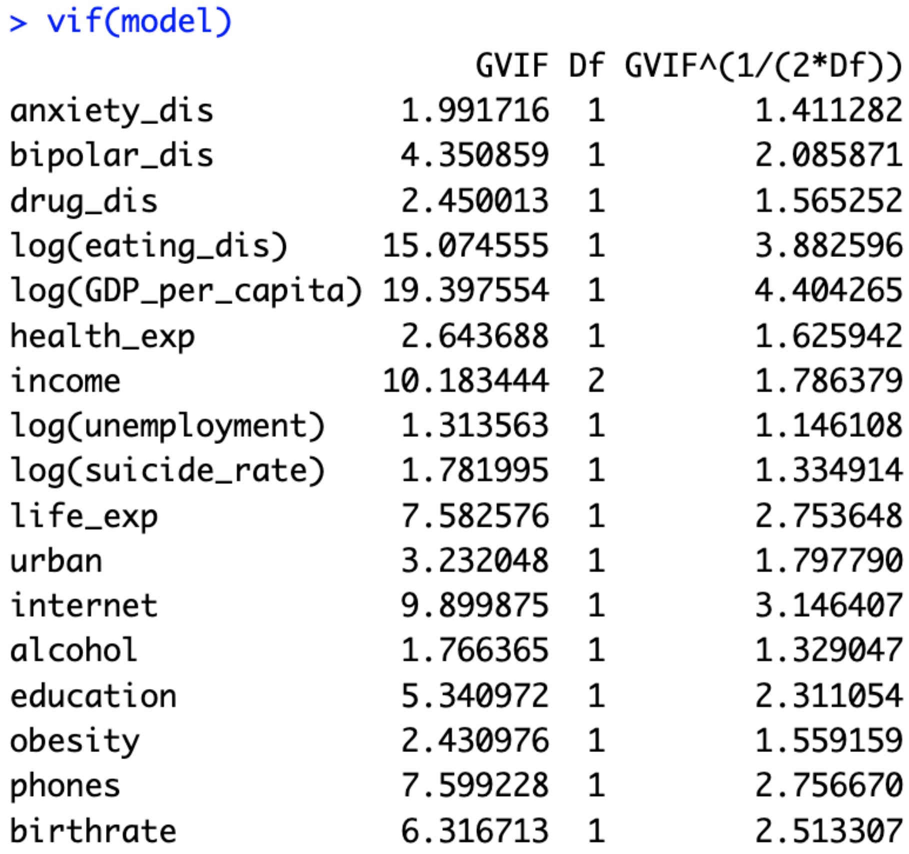


Figure 29: Histogram of residuals

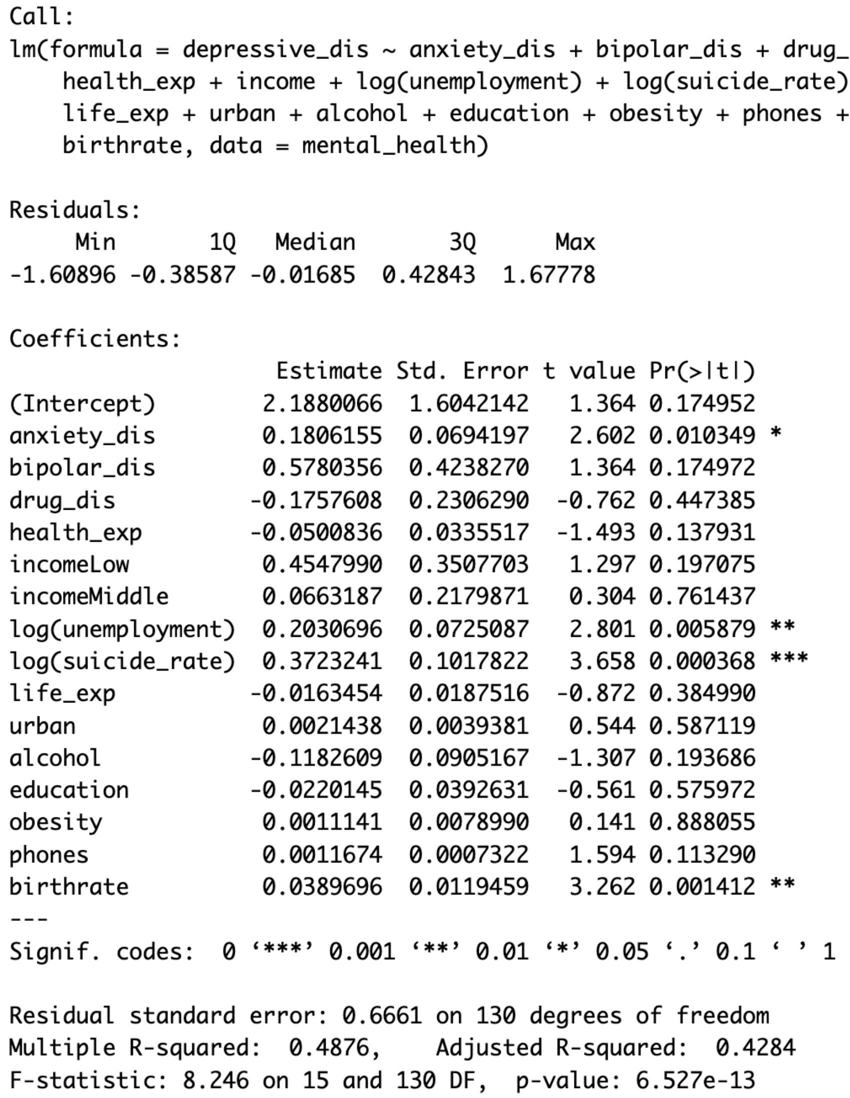
The last assumption we have to move beyond is that of multicollinearity. Multicollinearity arises when an independent variable is a linear function of other independent variables. Therefore, the next step is to eliminate variables with a very high multicollinearity value from the final model. To undertake this path, it is necessary to calculate the VIF (Variance Inflation Factor). This value indicates how much the estimate of one regression coefficient is influenced by the estimates of the other regression coefficients. In addition, the output has another value, namely the GVIF$\hat{(}$1/(2*Df)). This figure indicates the inflation factor of the adjusted generalised error and is easier to interpret. Therefore, we will eliminate variables that have this last value (the third column) greater than 3. The outcome is presented below.



```
> 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
```

Figure 30: VIF results

As can be seen from the output, it can be stated that for the reasons just mentioned, the following variables will not be used for the final regression model: 'log(eating dis)', 'log(GDP per capita)', 'internet'. The first multiple linear regression model, having eliminated all outliers and having coped with multicollinearity and distribution problems obtained, is shown below.



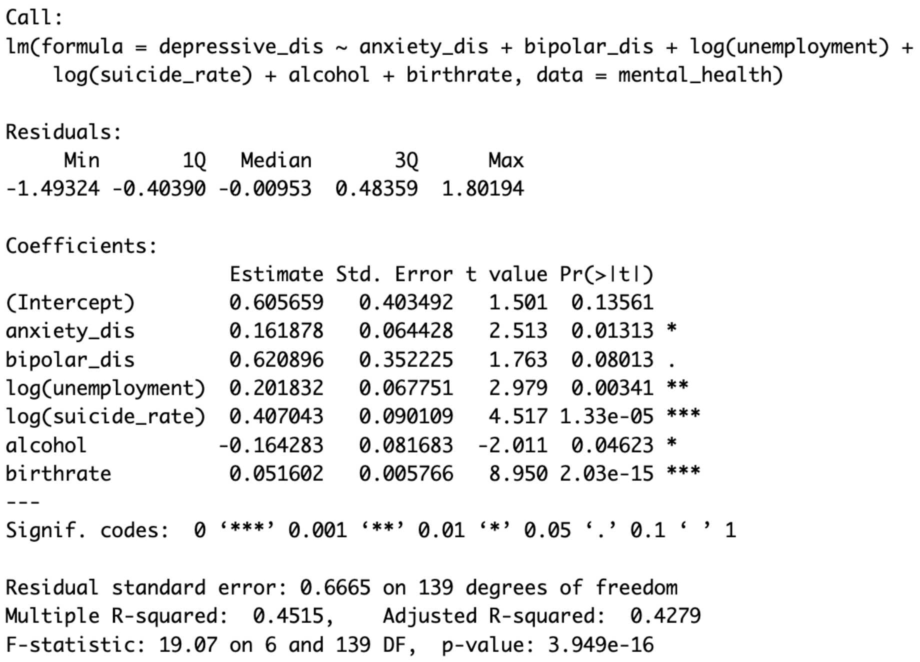
```
Call:
lm(formula = depressive_dis ~ anxiety_dis + bipolar_dis + drug_dis +
    health_exp + income + log(unemployment) + log(suicide_rate) +
    life_exp + urban + alcohol + education + obesity + phones +
    birthrate, data = mental_health)

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
```

Now, through an iterative process of eliminating the variables that have the largest p-value, we obtain the final manually run regression model. This process allows us to identify the variables that contribute significantly to the prediction of the dependent variable.



```
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
```

Looking at the output, a few conclusions can be drawn. First of all, one can talk about the coefficients. An increase of one unit in the estimated percentage of people with anxiety disorders is associated with an estimated increase of 0.162 in the percentage of people with depression. This, suggests that anxiety may have a significant impact on depression. Although not statistically significant, a one-unit increase in the estimated percentage of people with bipolar disorder is associated with an estimated 0.621 increase in the percentage of people with depression. A one-unit increase in the natural logarithm of the unemployment rate is associated with an estimated increase of 0.202 in the percentage of persons with depression. In addition, an increase in the natural logarithm of the suicide rate is associated with a significant estimated increase in the percentage of people with depression. This is consistent with the known association between suicide and depression. A one-unit increase in the estimated percentage of persons with an alcohol-related disorder is associated with an estimated decrease of 0.164 in the percentage of persons with depression. This value may be counterintuitive. Finally, a one-unit increase in the birth rate is associated with an estimated increase of 0.052 in the percentage of persons with depression. Finally, the model has a multiple R-squared of 0.5415. This means that the model manages to capture approximately 45.15% of the observed variability in the dependent variable.

Finally, a graph showing the actual and predicted values of the dependent variable 'depressive dis' is presented. On the y-axis, we have the predicted values of the variable 'depressive dis' obtained from the linear regression model.

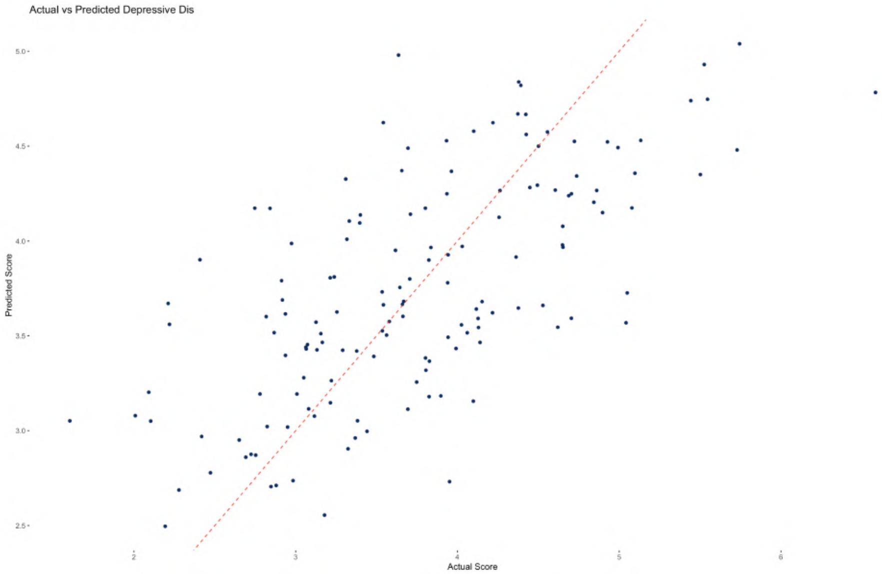
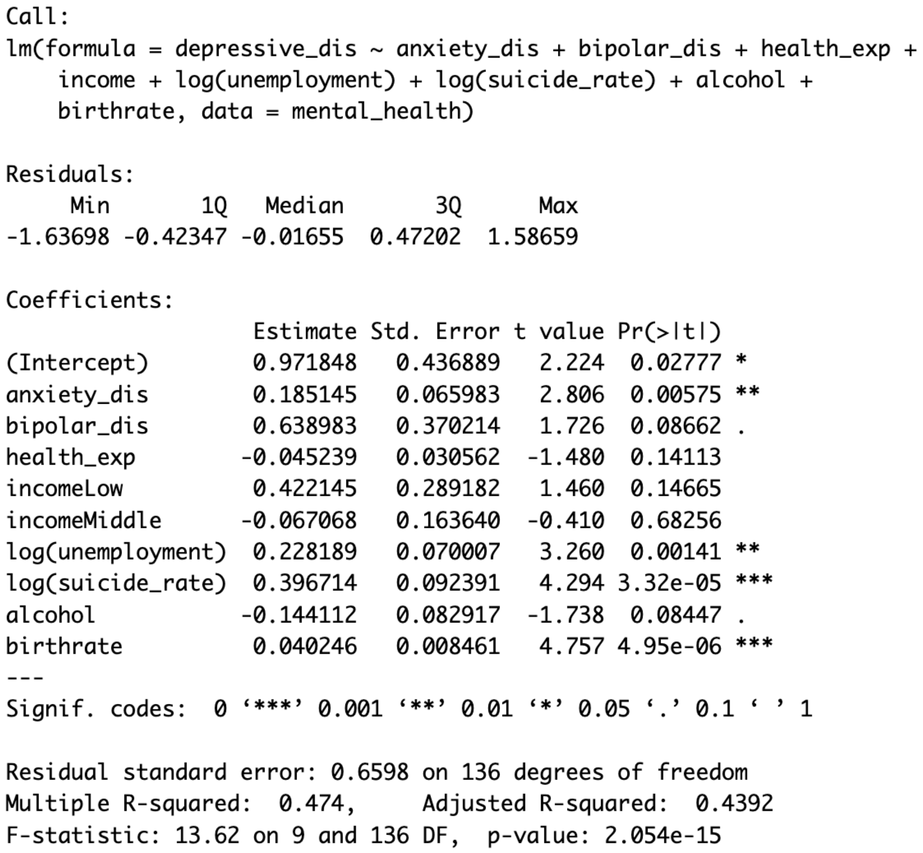


Figure 31: Actual versus predicted values

4.2 Stepwise Regression

After carrying out the regression manually by eliminating each variable according to the criterion of the largest p-value, the stepwise regression is performed. This automatic approach iteratively realises a regression model by selecting the independent variables. In the following regression model, the variables that contribute significantly to the model are selected according to Akaike's information criterion (AIC), which is a value that maximises the verisimilitude of the data. The output is shown below.



```
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
```

Figure 32: Stepwise regression

4.3 Ridge and Lasso Regression

Traditional Ordinary Least Squares (OLS) linear regression aims to minimize the sum of the squares of the errors between predicted and observed values. However, there are more sophisticated regression techniques that incorporate regularization and penalty to OLS coefficients. Two notable examples are Ridge and Lasso regression. These advanced regression methods constrain the complexity of the model by adjusting the coefficients of the variables. Ridge regression and Lasso regression are particularly effective in handling multicollinearity and preventing overfitting, making them valuable tools in predictive modeling and data analysis.

In Ridge regularisation, a penalty term proportional to the square of the absolute values of the model coefficients is added to keep the coefficients small but non-zero. This is done to make the model more stable and to prevent overfitting. In our case, the optimal regularisation parameter λ (lambda) is 0.1652566. The more λ tends to 0, the more the coefficients are the same as in linear regression. Conversely, the more λ increases, the more the coefficients tend towards 0. This can be observed from the graph below.

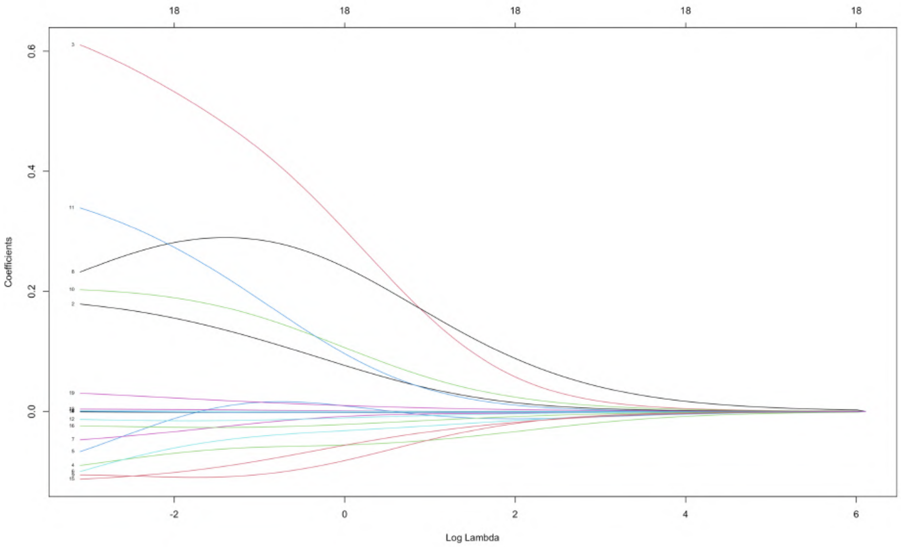
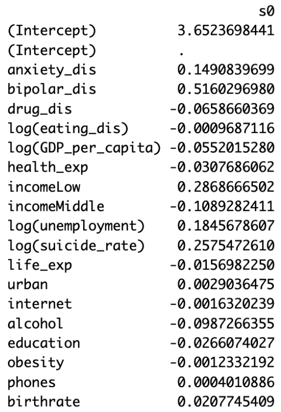


Figure 33: Ridge Regression



```
                               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
```

Figure 34: Ridge coefficients

For the Ridge regression, the coefficients are as shown below.

The coefficients represent the effect of each independent variable on the value of the dependent variable, taking into account the penalty added by the ridge regression method. This penalty allows for greater stability and interpretability of the model. This means that the coefficients obtained from ridge regression are generally smaller than in linear regression. Variables with non-zero coefficients are considered important by the model and have a significant impact on the dependent variable. Variables with coefficients close to zero may be considered less influential or have been penalised more by the ridge regression method.

In contrast, a penalty term proportional to the absolute value of the model's coefficients is added for Lasso regression. Unlike Ridge regularisation, Lasso regularisation can cause some coefficients to become exactly zero, making the model more interpretable and automatically selecting a subset of the variables. In our case, the value of λ is 0.03644125. This value is relatively low, thus having coefficients subject to strong regularisation. Thus, the coefficients tend to be reduced to 0 compared to the Ridge. The graph to observe λ is as follows.

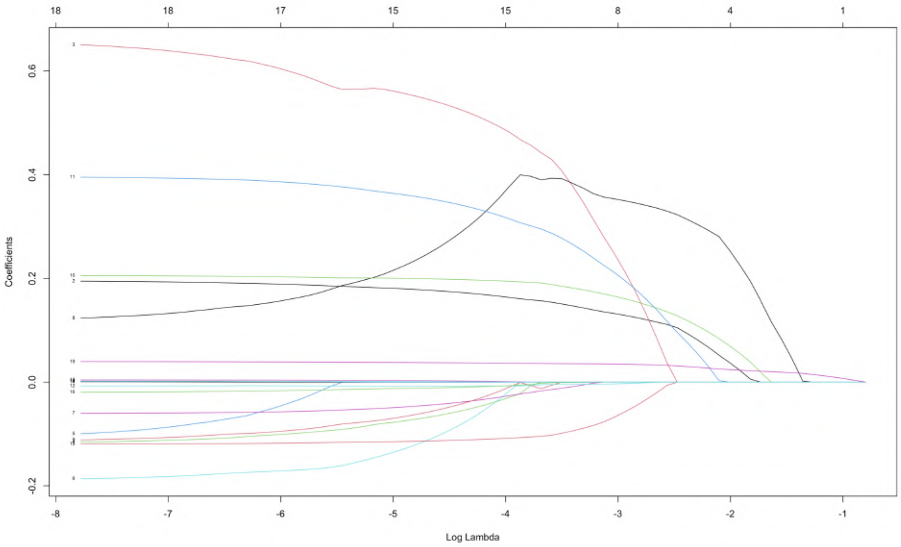
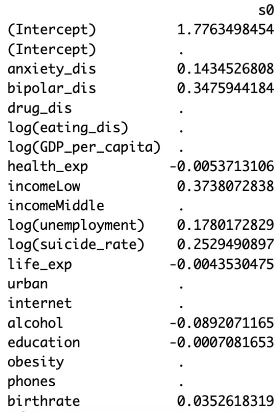


Figure 35: Lasso regression



```
                              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
```

Figure 36: Lasso coefficients

As can be seen from the output on the coefficients of the Lasso regression, some of them are reduced to zero. This means that these variables did not contribute significantly to the prediction of the variable on the percentage of persons suffering from the disease of depression. The variables that reduced to zero are 'drug dis', 'log(eating dis)', 'log(GDP per capita)', 'urban', 'internet', 'obesity', 'phones'. The other variables are those that the model considers relevant to explain the dependent variable. To give an example, a one-unit increase in people with anxiety is associated with a 0.143 increase in the output of the model, holding all other variables fixed.

To evaluate which of the two models performs better, it is essential to compare their performance using the Mean Quadratic Error (MSE) metric. The MSE of the ridge regression is 0.4084427, while that of Lasso is 0.4221251. Therefore, since the MSE of the ridge regression is lower than that of Lasso, we can conclude that the ridge regression model has a better predictive performance than Lasso in this context. Indeed, a lower MSE value indicates a better fit of the model to the training data and, consequently, a better predictive performance of the model.

In addition, the R-squared value of the Ridge regression is 0.4763796, while that of the Lasso regression is 0.4583483. the Ridge regression model has a slightly higher R-squared value than the Lasso model, indicating that the Ridge regression model is able to explain a greater amount of variation in the observed data than the Lasso model.

4.4 Decision Tree

The objective of this methodology is to obtain a hierarchical segmentation of a set of units, sometimes very large, through the identification of 'rules' that exploit the relationship existing between the class to which they belong and the variables detected for each unit. The graphic output of the procedure consists of a tree structure - with nodes, branches and leaves - that has points of contact with the dendrogram of cluster analysis. The application of decision trees, on the other hand, requires the a priori knowledge of the class to which each unit belongs: the objective of the technique is in fact to identify the optimal decision rule, i.e. the rule that, given a certain set of surveyed variables, best predicts the class to which each unit belongs.

The output obtained is intended to predict the continuous dependent variable, the depression index. In fact, continuous variable decision trees are used to make predictions. The data are broken down according to certain criteria. The graph consists of nodes and leaves. Nodes are the points at which the data are split and are associated with an input value. Leaves, on the other hand, are the intermediate or final results and are associated with an output value. The leaves show the data once it has been split.

Furthermore, the mental health dataset was divided into two datasets, training and test, using a specified proportion (70% for training and 30% for test) as a criterion. Therefore, using the *sample()* function, observations were selected randomly and we obtained that the training dataset consisted of 103 observations and the test dataset of 43. We then went on to train a decision tree model using the training dataset using as the dependent variable, of course, the percentage of people in a nation suffering from the disease of depression. Finally, in order to understand the decision rules and significant variables for predicting 'depressive dis', a graphical representation of the decision tree is shown with its nodes representing the variables selected by the algorithm.

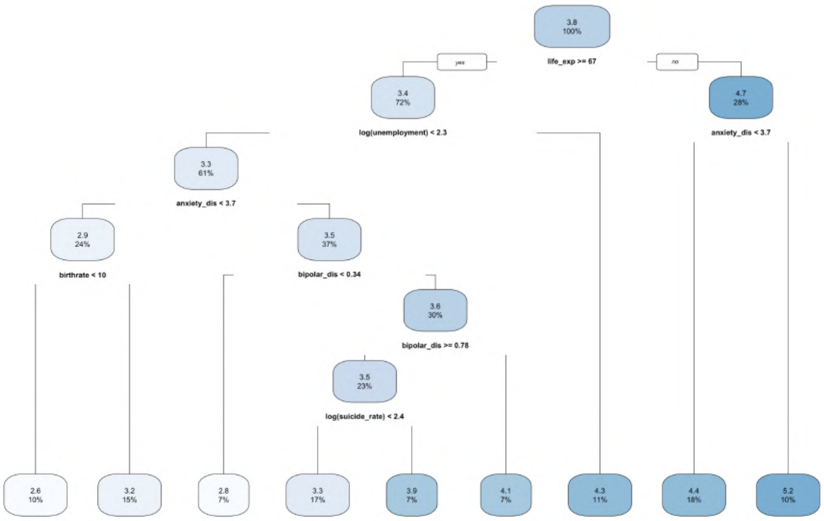


Figure 37: Decision Tree

As can be seen from the graph, the variables selected to explain the model are the following: anxiety dis, bipolar dis, birthrate, life exp, log(suicide rate) and log(unemployment). Each leaf of the tree represents the predictive value of the response variable, that is, the percentage of depressed people in a country, together with the percentage of total observations in the dataset that fall within that specific leaf and fulfil the specified conditions. In the analysis of the graph, we notice that on the left are the countries with a lower incidence of depression. In this group, we see around 10 countries with a life expectancy of at least 67 years, an unemployment rate of less than 0.83, a percentage of people with anxiety of less than 3.7% and a birth rate of less than 10 births per 1000 inhabitants. On average, 2.6% of these people suffer from depression. It is possible to assess each leaf on a case-by-case basis. On the right-hand side, however, we find about 10 countries with a life expectancy of less than 67 years, more than 3.7% of people with anxiety and, on average, a 'depressive dis' score of 5.2.

Analysing the performance of the decision tree-based predictive model against the training data, we observe that the output obtained indicates that the decision tree-based predictive model has an RMSE (Root Mean Square Error) of approximately 0.5204 and an R-squared of approximately 0.6933 when tested on the training data itself. The RMSE measures the accuracy of the model's predictions, while the R-squared represents the percentage of variation in the dependent variable that is explained by the independent variables in the model. A lower RMSE and higher R-squared indicate a higher accuracy and fit of the model compared to the training data. Furthermore, the mean square error is 0.270865.

5 Conclusion

This study focused on an in-depth analysis of an extremely topical issue, which is mental health, with a particular focus on depression. This widespread mood disorder can result from a variety of factors, which can vary from individual to individual and also according to the cultural, economic and social aspects of a nation. The fundamental aim of this report was to investigate the determinants of depression and identify which blocks of states are most affected by this problem.

As highlighted in the initial phase of the research, a significant prevalence of depression emerges in African countries (although in thirteenth place in our ranking we find a European country such as Greece), mainly attributable to unfavourable socio-economic conditions. However, it should be emphasised that the mental health problem is not exclusively confined to less developed nations. Even in more advanced countries, anxiety, eating disorders, bipolarism and suicide rates are major issues. For example, Portugal has a high percentage of its population affected by anxiety, while Australia has a worrying incidence of eating disorders and New Zealand ranks high in bipolarism. Moreover, it is necessary to affirm that East and South-East Asian countries, such as Brunei, Singapore, South Korea, Japan and Myanmar, were found to show a low incidence of depression. These states, with their unique cultural, social and economic dynamics, seem to be associated with a lower prevalence of this mood disorder.

Through unsupervised analysis, we examined the behaviour of the various variables within each country. By combining k-means clustering and hierarchical clustering, we identified two distinct groups of nations that emerge by processing the two algorithms. The first group includes nations from Western Europe, North America, Australia and New Zealand, while the second group includes countries from Central Africa and South-East Asia, such as India, Pakistan and Afghanistan. The nations in the first group are characterised by a high level of economic, technological and educational development, along with significant investments in the health sector. However, they also have a significant prevalence of mental health disorders, such as anxiety, bipolarism, eating disorders, alcoholism, drug abuse and high suicide rates. This shows that economic prosperity does not necessarily guarantee good mental health. In the second group, on the other hand, there is a high incidence of depression and a particularly high birth rate. We have also identified other groups of nations that share many similarities. For example, the countries of South and Central America show common characteristics, as do the countries of Central Asia. Russia and Eastern European nations also show many similarities, suggesting a heritage of Russian influence in these regions.

In the supervised phase, we aimed to identify the variables that most influence the rate of depression in a country. Through several regression analyses, six relevant variables emerged: anxiety, bipolarism, unemployment, suicide rate, alcohol consumption and birth rate. It must be stated, however, that there is an unexpectedly slight negative correlation between alcohol abuse and the level of depression. Additionally, the Lasso regression included public health investment and education level as significant factors. The decision tree also highlighted the importance of life expectancy. These findings indicate that depression is influenced by social factors such as anxiety, bipolar disorder, suicide, alcoholism and birth rate, by economic factors such as unemployment and healthcare spending, and by cultural factors such as education level.

6 Sitography

The World Bank

- GDP pre-capita - https://data.worldbank.org/indicator/NY.GDP.PCAP.CD?end=2022&start=1960&view=chart
- Unemployment - https://data.worldbank.org/indicator/SL.UEM.TOTL.ZS?end=2022&start=1991&view=chart
- Current health expenditure (% of GDP) - https://data.worldbank.org/indicator/SH.XPD.CHEX.GD.ZS
- Population, total - https://data.worldbank.org/indicator/SP.POP.TOTL
- Life expectancy - https://data.worldbank.org/indicator/SP.DYN.LE00.IN?end=2019&most_recent_year_desc=true&start=2019
- Urban population - https://data.worldbank.org/indicator/SP.URB.TOTL.IN.ZS?end=2019&start=2019
- Internet access - https://data.worldbank.org/indicator/IT.NET.USER.ZS

Our World in Data

- Depression - https://ourworldindata.org/grapher/depressive-disorders-prevalence-ihme
- Bipolar disorder - https://ourworldindata.org/grapher/anxiety-disorders-prevalence
- Eating disorder - https://ourworldindata.org/grapher/eating-disorders-prevalence
- Anxiety disorder - https://ourworldindata.org/grapher/anxiety-disorders-prevalence
- People with drug use disorders - https://ourworldindata.org/grapher/number-with-drug-use-disorders-country
- Suicide Rate - https://ourworldindata.org/grapher/death-rate-from-suicides-ghe
- Alcohol disorder - https://ourworldindata.org/grapher/share-with-alcohol-use-disorders

Global Health Expenditure Database

- Income - https://apps.who.int/nha/database

Human Development Reports

- Average years of schooling - `https://hdr.undp.org/data-center/human-development-index#/indicies/HDI`

World Health Organization

- Obesity - `https://data.who.int/indicators/i/BEFA58B`