Multidimensional Data Visualization

Antonio De Patto

May 7, 2025

Abstract

In this paper we will examine the structure of a multidimensional dataset by conducting an in-depth analysis of its features and objects. The goal will be to understand the meaning of each variable in order to reduce dimensionality and avoid redundancy. At the end of the paper we will have a complete usable dataset ready for any type of analysis.

1 The dataset

The dataset we will analyze consists of 195 rows and 35 columns, with each row representing a different country. Various indicators are associated with each nation, providing a comprehensive overview of its economic, demographic, environmental, and health aspects. The following tables present the data set in its entirety without having performed any transformation of the data. We can notice from the beginning the presence of columns that do not contribute anything to our analysis since they are not economic indicator. These include the following attributes:

- Abbreviation: abbreviation or code representing the country
- Calling Code: international calling code for the country
- Capital: name of the capital or major city
- Currency-Code: currency code used in the country
- Largest city: name of the country's largest city
- Language: official language(s) spoken in the country
- Latitude: latitude coordinate of the country's location
- Longitude: longitude coordinate of the country's location

	Country	Density(P/Km2)	Abbreviation	Agricultural Land(%)	Land Area(Km2)	Armed Forces	Birth Rate
0	Afghanistan	60	AF	58.10%	652,230	323,000	32.49
1	Albania	105	AL	43.10%	28,748	9,000	11.78
2	Algeria	18	DZ	17.40%	2,381,741	317,000	24.28
3	Andorra	164	AD	40.00%	468	NaN	7.20
4	Angola	26	AO	47.50%	1,246,700	117,000	40.73
...	...	...	...	...	...	...	...
190	Venezuela	32	VE	24.50%	912,050	343,000	17.88
191	Vietnam	314	VN	39.30%	331,210	522,000	16.75
192	Yemen	56	YE	44.60%	527,968	40,000	30.45
193	Zambia	25	ZM	32.10%	752,618	16,000	36.19
194	Zimbabwe	38	ZW	41.90%	390,757	51,000	30.68

Figure 1: The table above indicates the first 7 columns

	Calling Code	Capital	Co2-Emissions	CPI	CPI Change(%)	Currency-Code	Fertility Rate
0	93.0	Kabul	8,672	149.9	2.30%	AFN	4.47
1	355.0	Tirana	4,536	119.05	1.40%	ALL	1.62
2	213.0	Algiers	150,006	151.36	2.00%	DZD	3.02
3	376.0	Andorra la Vella	469	NaN	NaN	EUR	1.27
4	244.0	Luanda	34,693	261.73	17.10%	AOA	5.52
...	...	...	...	...	...	...	...
190	58.0	Caracas	164,175	2,740.27	254.90%	VED	2.27
191	84.0	Hanoi	192,668	163.52	2.80%	VND	2.05
192	967.0	Sanaa	10,609	157.58	8.10%	YER	3.79
193	260.0	Lusaka	5,141	212.31	9.20%	ZMW	4.63
194	263.0	Harare	10,983	105.51	0.90%	NaN	3.62

Figure 2: The table above indicates columns from the 8th to the 14th

	Life expectancy	Maternal mortality ratio	Minimum wage	Language	Self-paid Health	Doctors/1000	Population
0	64.5	638.0	$0.43	Pashto	78.40%	0.28	38,041,754
1	78.5	15.0	$1.12	Albanian	56.90%	1.20	2,854,191
2	76.7	112.0	$0.95	Arabic	28.10%	1.72	43,053,054
3	NaN	NaN	$6.63	Catalan	36.40%	3.33	77,142
4	60.8	241.0	$0.71	Portuguese	33.40%	0.21	31,825,295
...	...	...	...	...	...	...	...
190	72.1	125.0	$0.01	Spanish	45.80%	1.92	28,515,829
191	75.3	43.0	$0.73	Vietnamese	43.50%	0.82	96,462,106
192	66.1	164.0	NaN	Arabic	81.00%	0.31	29,161,922
193	63.5	213.0	$0.24	English	27.50%	1.19	17,861,030
194	61.2	458.0	NaN	Shona	25.80%	0.21	14,645,468

Figure 3: The table above indicates columns from the 15th to the 21st

	Forested Area(%)	Gasoline Price	GDP	Primary educ. enr.(%)	Tertiary educ. enr.(%)	Infant mortality	Largest city
0	2.10%	$0.70	$19,101,353,833	104.00%	9.70%	47.9	Kabul
1	28.10%	$1.36	$15,278,077,447	107.00%	55.00%	7.8	Tirana
2	0.80%	$0.28	$169,988,236,398	109.90%	51.40%	20.1	Algiers
3	34.00%	$1.51	$3,154,057,987	106.40%	NaN	2.7	Andorra la Vella
4	46.30%	$0.97	$94,635,415,870	113.50%	9.30%	51.6	Luanda
...	...	...	...	...	...	...	...
190	52.70%	$0.00	$482,359,318,768	97.20%	79.30%	21.4	Caracas
191	48.10%	$0.80	$261,921,244,843	110.60%	28.50%	16.5	Ho Chi Minh City
192	1.00%	$0.92	$26,914,402,224	93.60%	10.20%	42.9	Sanaa
193	65.20%	$1.40	$23,064,722,446	98.70%	4.10%	40.4	Lusaka
194	35.50%	$1.34	$21,440,758,800	109.90%	10.00%	33.9	Harare

Figure 4: The table above indicates the columns of the 22nd to the 28th

	Labor force participation(%)	Tax revenue(%)	Total tax rate	Unemployment rate	Urban population	Latitude	Longitude
0	48.90%	9.30%	71.40%	11.12%	9,797,273	33.939110	67.709953
1	55.70%	18.60%	36.60%	12.33%	1,747,593	41.153332	20.168331
2	41.20%	37.20%	66.10%	11.70%	31,510,100	28.033886	1.659626
3	NaN	NaN	NaN	NaN	67,873	42.506285	1.521801
4	77.50%	9.20%	49.10%	6.89%	21,061,025	-11.202692	17.873887
...	...	...	...	...	...	...	...
190	59.70%	NaN	73.30%	8.80%	25,162,368	6.423750	-66.589730
191	77.40%	19.10%	37.60%	2.01%	35,332,140	14.058324	108.277199
192	38.00%	NaN	26.60%	12.91%	10,869,523	15.552727	48.516388
193	74.60%	16.20%	15.60%	11.43%	7,871,713	-13.133897	27.849332
194	83.10%	20.70%	31.60%	4.95%	4,717,305	-19.015438	29.154857

Figure 5: The table above indicates the columns of the 29th to the 35th

By eliminating variables that don't add value to our analysis, we can perform an initial screening and proceed with our preliminary analyses. The attributes have been reordered according to their corresponding macroeconomic sectors, as can be seen from table 6. The first three variables are categorized as 'demographic', while variables 4 through 12 — ranging from 'Armed Forces' to 'CPI Change (%)'— are purely economic. The variables from 'Doctors/1000' up to the 19th are related to the health sector. Variables 20 through 24 are linked to the environment and sustainability, and the final two variables belong to education and instruction, completing the dataset.

#	Column	Non-Null Count	Dtype
0	Country	195 non-null	object
1	Density(P/Km2)	195 non-null	float64
2	Urban population	190 non-null	float64
3	Population	194 non-null	float64
4	Armed Forces	171 non-null	float64
5	GDP	193 non-null	float64
6	Minimum wage	150 non-null	float64
7	Labor force participation(%)	176 non-null	float64
8	Unemployment rate	176 non-null	float64
9	Tax revenue(%)	169 non-null	float64
10	Total tax rate	183 non-null	float64
11	CPI	178 non-null	float64
12	CPI Change(%)	179 non-null	float64
13	Doctors/1000	188 non-null	float64
14	Self-paid Health	188 non-null	float64
15	Birth Rate	189 non-null	float64
16	Fertility Rate	188 non-null	float64
17	Life expectancy	187 non-null	float64
18	Infant mortality	189 non-null	float64
19	Maternal mortality ratio	181 non-null	float64
20	Co2-Emissions	188 non-null	float64
21	Land Area(Km2)	194 non-null	float64
22	Forested Area(%)	188 non-null	float64
23	Agricultural Land(%)	188 non-null	float64
24	Gasoline Price	175 non-null	float64
25	Primary educ. enr.(%)	188 non-null	float64
26	Tertiary educ. enr.(%)	183 non-null	float64

Figure 6: Missing values for each variable

Going more deeply into the analysis, it can be seen that most variables contain missing values, but only in small percentages. As a result, when we move forward with selecting variables and handling missing data in the next chapter, we will still retain a sufficiently large dataset for our analyses. Furthermore, all the variables that are present in the data set are of numeric type except for the categorical variable relating to countries.

In the next chapter we are going to reduce the dimensionality and remove the missing values in order to finalize our dataset. By doing that we are going to make our analyses less redundant and much more meaningful, as well as making comparisons between countries much easier to interpret.

1.1 The choice of variables

The choice of variables to be retained to initially reduce the dimensionality of the dataset will be made on the basis of the correlation matrix. The variables that will be kept must also explain every economic aspect of a nation in its entirety. Therefore, the chosen indices will cover key dimensions, including demographic, economic, health, and educational factors, ensuring a well-rounded analysis.

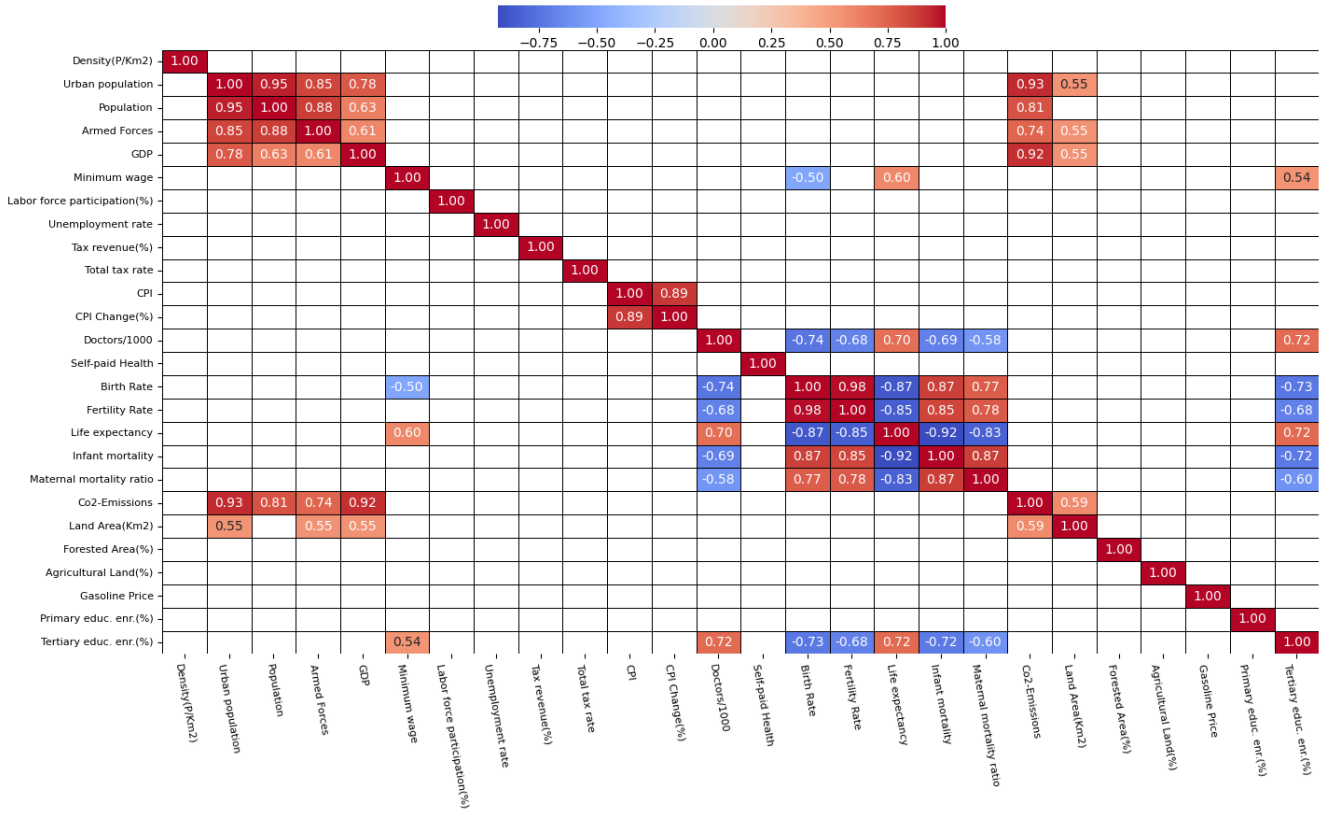


Figure 7: Correlation matrix

Having a large number of variables to interpret and to improve the readability of the correlation matrix we will represent only the values in which the correlation is greater than 0.5. We now present the variables that I personally choose to keep within the data set, trying to explain in an exhaustive way the reasons for my choices.

1.1.1 Demography attributes

- **Density (P/km2)**: starting from the variable 'Density (P/km2)' which represents the population density for each square kilometer we can say that this does not seem to be correlated with any other variable. It also represents a variable of great value for understanding the economic and demographic wealth of a country. Very often highly populated countries are indicators of progress and economic wealth, which attract immigration flows. It should also be said, however, that too high a population density can be associated with overcrowding and environmental problems that lead to greater pressure on infrastructure, public services and natural resources. For this reason, this variable should be analyzed together with other indicators such as GDP

in order to have a more detailed overview of a country. Other demographic variables could be included in the data set. An example is the variable 'Population' which however turns out to be less economically significant than the variable 'Density'. This is because the variable 'Density' takes into account the differences in size of the countries making the comparisons much more fair. The variable 'Population' on the other hand provides only a raw number, which can be misleading when comparing different-sized regions. The variable 'Urban population' has a greater value because it indicates the percentage of people living in inhabited centers. However, since it is strongly correlated with the variables 'Armed Forces' or 'GDP' it is redundant to include it among our variables and create correlation problems if we want to use the data set to perform predictive regressions

1.1.2 Economic attributes

- **Armed Forces**: the variable 'Armed Forces' measures the war potential of a country and is an indirect index of how much GDP is dedicated by each state to personal defense. The variable is measured in thousands and indicates the total number of people who serve in the army. A high number of military members may reflect high defense spending and a strong focus on national security. Furthermore, countries with a high level of public spending dedicated to national defense play a strategic role at the international level, influencing political balances and global alliances. As can be seen in Figure 7, this variable is correlated with the demographic variables 'Population' and 'Urban population', which however were not considered in our analyses. The correlation is also present with GDP, although to a lesser extent. The variable is correlated more strongly with 'Co2-Emissions', which however will not be taken into consideration because it is strongly correlated with the variable 'GDP'

- **GDP**: this variable is a very famous economic index that represents the Gross Domestic Product, which would be the total value of goods and services produced in the country. It is a fundamental variable to indicate the economic importance of a country and therefore it will be included in our data set. A high GDP generally indicates a developed and dynamic economy with better infrastructure, public services and job opportunities.

- **Labor force participation(%)**: this variable indicates the percentage of the active population participating in the labor market, including both employed and unemployed people looking for work. A low value may indicate barriers to entry into the labor market, such as gender discrimination or lack of opportunities for young people. A high participation indicates an economy with an active and productive labor force, while a low value may reflect structural problems or a high number of unemployed people (students, retired people, housewives, etc.). Comparing this variable with GDP, unemployment rate and education level can help to better understand the economic state of a country.

- **Unemployment rate**: in the same way we include this very important economic index that represents the percentage of the workforce that is unemployed but actively seeking employment. The unemployment rate is in fact a direct measure of the economic prosperity of a country as a high unemployment rate may indicate economic hardship, while a low rate suggests a solid labor market.

- **Total tax rate**: this variable related to the 'taxation' sector is inserted to measures the overall burden of taxation on businesses, expressed as a percentage of trade profits. It includes corporate income taxes, labor taxes, and other mandatory contributions. It is an important economic indicator because a high tax rate can reduce the attractiveness of a country for investment, while a lower one can encourage entrepreneurship, and therefore a high tax burden can influence business growth, job creation, and innovation.

- **CPI**: this variable measures the price development of a representative basket of goods and services consumed by households. It is a key tool for assessing inflation and purchasing power in a country. An increase in the CPI indicates rising prices, signaling inflationary pressures in the economy while a value that is too low (or negative) may suggest deflation, with possible negative consequences for the economy. It is a key parameter for central banks' decisions on interest rates and anti-inflation measures. It is therefore of great importance to include this variable in our dataset not only for its economic value but also because it does not appear to be highly correlated with any other variable.

1.1.3 Health and Insurance attributes

- **Self-paid Health**:this variable is considered in our dataset because it represents a good indicator of the health care provided by a country to its citizens. Formally, this attribute indicates the percentage of health

care expenditure that is paid directly by citizens, without coverage by the state or insurance. In this case, a high percentage indicates that the health care system is highly private or that public coverage is insufficient, making care less accessible for the weakest segments of the population, leading in some cases to forgoing medical care. A low value suggests that the government or insurance covers most of the costs, ensuring greater equity in access to health services. This variable was also chosen for its low correlation with the other variables, and therefore constitutes a statistically significant indicator of a country's health sector.

- **Birth Rate**:represents the number of births per 1,000 inhabitants in a year. It is a fundamental indicator for understanding population growth and the socioeconomic dynamics of a country. This value is of great importance because it affects the future availability of labor force and the ratio between the active and inactive population (young and old), influencing the planning of essential services such as schools, health care and pensions. This variable is highly correlated with the variables 'Fertility Rate', 'Infant mortality' and 'Maternal Mortality ratio' while it presents a negative correlation with the variables 'Life expectancy' and 'Doctor/1000'. Obviously it would be very useful to include these variables in our dataset but we would run into problems of redundancy and multicollinearity and therefore we proceed to keep only 'Birth Rate' as a representative variable of this macro group.

1.1.4 Environment and Sustainability attributes

- **Forested Area(%)**: this attribute represents the percentage of a country's territory covered by forests. This indicator is crucial to assess the environmental health and sustainability of a nation's natural resources. A high percentage of forest area is indicative of rich biodiversity and natural conservation policies as well as being a resource for the economy, through wood and other natural resources. Looking at the correlation matrix in figure 7 this variable does not appear to be correlated with any other variable.

- **Agricultural Land(%)**:represents the percentage of a country's territory used for agriculture, which includes lands used for crops, pastures and other agricultural activities. This indicator is fundamental to understanding the role of agriculture in the economy and land management. A high percentage of agricultural land may suggest a strong dependence on the agricultural sector for the economy and food security of a country. However, high agricultural use could lead to deforestation or unsustainable practices. Therefore, when considered together with the variable 'Forested Area', the sustainability of land use and the compatibility of agricultural policies with environmental protection can be assessed.

1.1.5 Education and Instruction attributes

- **Primary educ. enr.(%)**: represents the gross enrollment ratio for primary education, which is the percentage of children enrolled in primary school compared to the population of primary school age. This figure can exceed 100% when children are enrolled who are older or younger than the typical primary school age. A high enrollment ratio suggests that most children have access to primary education, a key indicator of a country's social and economic progress. It is therefore a crucial indicator for analyzing the level of human development and the ability of a country to ensure equal educational opportunities for all its citizens. This variable was also included because it is not highly correlated with any other variable, unlike the attribute 'Tertiary educ. enr.'. This represents the gross enrollment rate in tertiary education, i.e. the percentage of the university-age population enrolling in higher education, including universities and other institutions of advanced learning. However, since it is highly correlated with variables already present in our dataset, it makes sense not to include it.

1.2 Selection of variables

We thus obtain a dataset that contains 12 of the 35 initial variables and 160 observations, 35 less than the initial number of observations. However, the dataset now has great interpretability and is free of missing values, which makes it usable for analysis. Figure 9 presents the correlation matrix of the final dataset, which as we can see presents independent variables that can be used to perform regressions and forecasts.

```
Data columns (total 13 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   Country                       160 non-null    object
 1   Density(P/Km2)                160 non-null    float64
 2   Armed Forces                  160 non-null    float64
 3   GDP                           160 non-null    float64
 4   Labor force participation(%)  160 non-null    float64
 5   Unemployment rate             160 non-null    float64
 6   Total tax rate                160 non-null    float64
 7   CPI                           160 non-null    float64
 8   Self-paid Health              160 non-null    float64
 9   Birth Rate                    160 non-null    float64
 10  Forested Area(%)              160 non-null    float64
 11  Agricultural Land(%)          160 non-null    float64
 12  Primary educ. enr.(%)         160 non-null    float64
```

Figure 8: Final data set composition

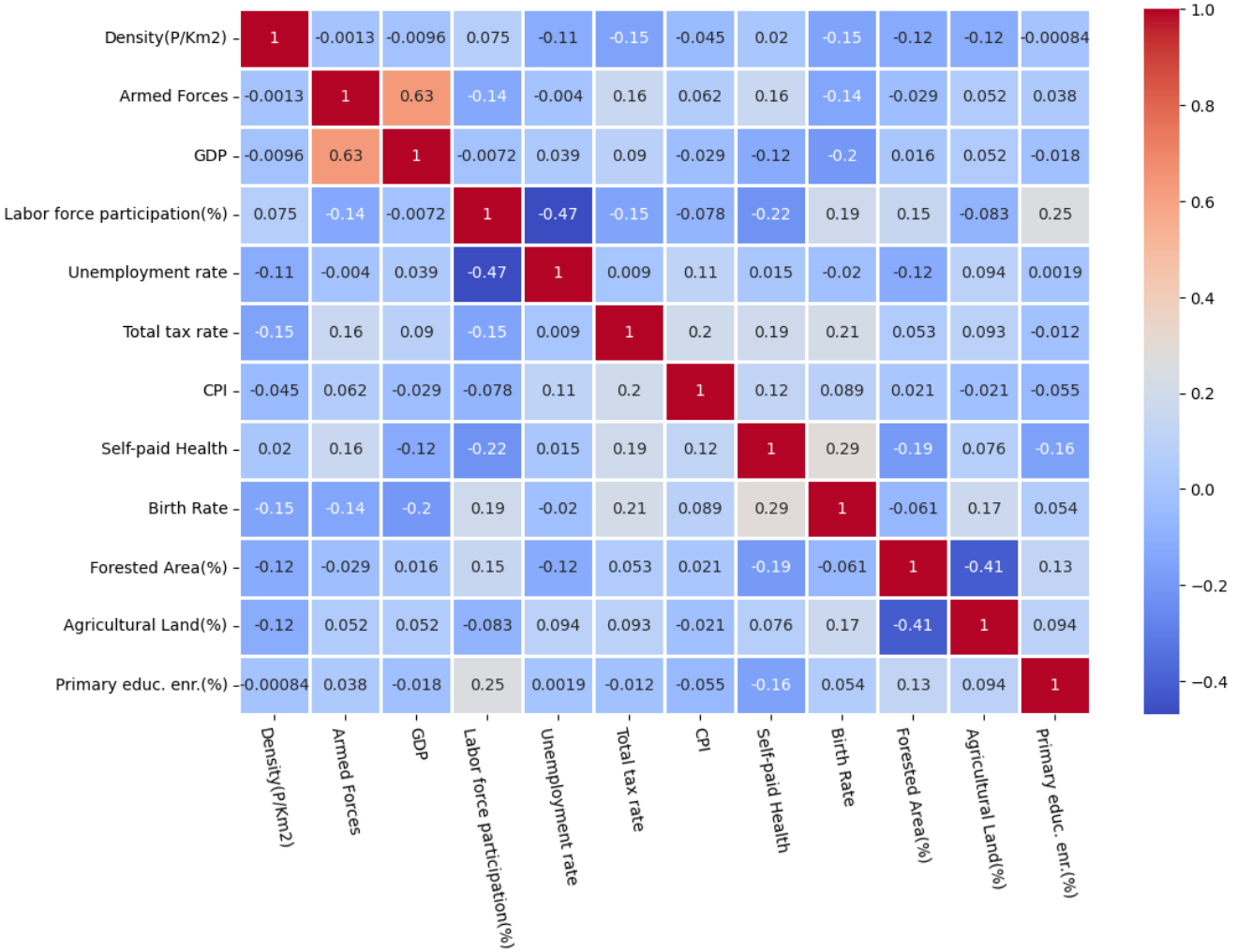


Figure 9: Correlation matrix of the final data set

Looking at figure 10, one can get a clearer idea of the composition of each variable, as descriptive statistics such

as mean, standard deviation, minimum and maximum value, and quartiles at 25, 50, and 75 percent are presented. The analysis of the dataset highlights significant differences between the countries represented. The population density (Density P/Km^2) varies enormously, with a minimum of just 2 inhabitants per km^2 and a maximum of 8,358. The average value stands at 212.76, but the high standard deviation (707.71) suggests a very heterogeneous distribution, with some countries extremely densely populated and others almost uninhabited.

The number of armed forces also shows strong variability: while some countries have no registered armed forces (minimum of 0), others have over 3 million military personnel. The average value is 156,300 units, with a standard deviation of 378,584.48, a sign that the data is influenced by some nations with very large armies.

Turning to the economy, GDP has an extremely wide range: the country with the lowest value has a GDP of around 429 million dollars, while the country with the highest GDP reaches 21.4 trillion. The average is around 573 billion dollars, but with a very high standard deviation (2.37 trillion), highlighting the great economic disparity between countries.

Employment is equally variable. The average labor force participation rate is 62.73%, with a variability between a minimum of 38% and a maximum of 86.8%. The unemployment rate, on the other hand, averages 6.8%, but with values ranging from 0.09% to 28.18%, a sign of strong differences in the employment situation between the various states.

From a fiscal point of view, the total taxation has an average value of 39.82%, with a minimum of 8% and a maximum of 106.3%, suggesting that some countries have much more onerous tax systems than others.

The consumer price index (CPI) shows extreme variability: the average value is 171.26, but the maximum exceeds 2,740, indicating strong inflation in some countries compared to others with more stable prices.

The healthcare system presents interesting differences: the share of healthcare expenditure borne by citizens(Self-paid Health) has an average of 33.47%, but can drop to 5.3% or rise to 81.6%, indicating very different financing models between countries.

From a demographic point of view, the birth rate varies from a minimum of 6.4 births per 1,000 inhabitants to a maximum of 46.08, with an average of 20.37, suggesting marked differences between countries with growing populations and others with very low birth rates.

With regard to the environment, forested land covers an average of 30.47% of the territory of the countries considered, but it ranges from nations completely devoid of forests to countries with a forest cover of 98.3%. Similarly, the share of agricultural land occupies an average of 39.46% of the territory, with variations between 0.6% and 82.6%, a sign of economies with very different levels of dependence on agriculture.

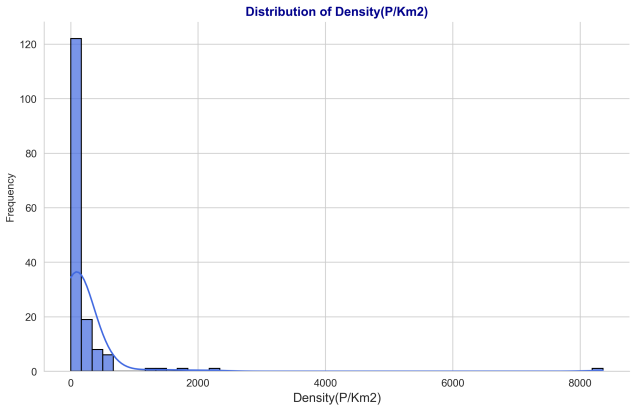
Finally, the rate of enrollment in primary education presents rather high values, with an average of 103.11% and a maximum of 142.5%. The presence of values above 100% can be explained by overlaps between different age groups or by the participation of adults in primary education programs.

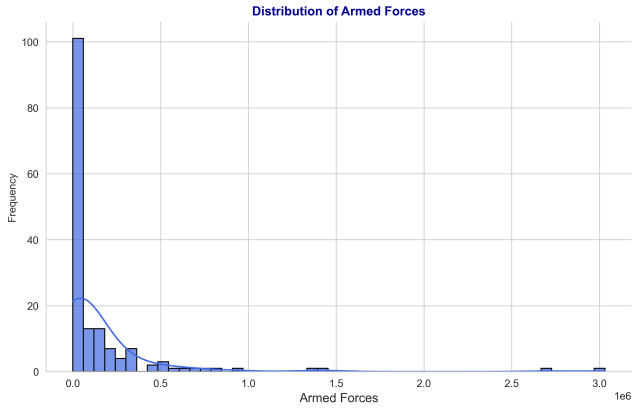
	Density(P/Km2)	Armed Forces	GDP	Labor force participation(%)	Unemployment rate	Total tax rate	CPI	Self-paid Health	Birth Rate	Forested Area(%)	Agricultural Land(%)	Primary educ. enr.(%)
count	160.00	160.00	1.600000e+02	160.00	160.00	160.00	160.00	160.00	160.00	160.00	160.00	160.00
mean	212.76	156300.00	5.733593e+11	62.73	6.80	39.82	171.26	33.47	20.37	30.47	39.46	103.11
std	707.71	378584.48	2.375542e+12	10.27	4.92	15.35	232.11	18.45	9.99	23.05	21.74	11.83
min	2.00	0.00	4.290166e+08	38.00	0.09	8.00	99.03	5.30	6.40	0.00	0.60	61.80
25%	31.75	10750.00	1.433814e+10	56.42	3.40	30.60	115.14	18.20	11.23	10.48	22.88	98.95
50%	83.00	29500.00	5.355460e+10	62.15	5.38	37.55	129.09	32.10	18.12	31.15	40.00	102.35
75%	152.25	136500.00	3.034813e+11	68.95	9.21	47.50	162.54	43.98	28.83	47.18	55.02	107.52
max	8358.00	3031000.00	2.142770e+13	86.80	28.18	106.30	2740.27	81.60	46.08	98.30	82.60	142.50

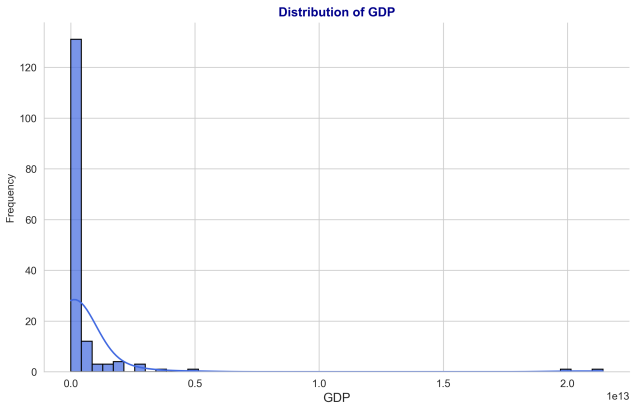
Figure 10: Summary statistics of the final data set

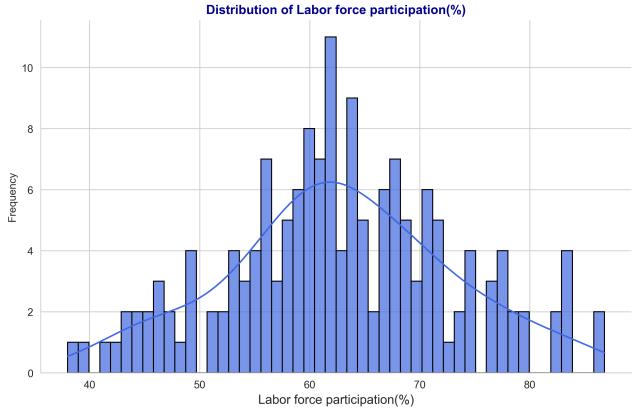
2 Data visualization

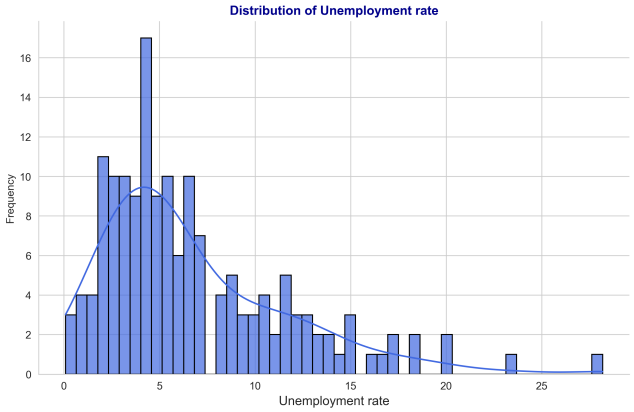
We therefore proceed with the first analyses of the dataset. The aim will be to extrapolate the meaning of the variables that make up the dataset and this will be done through different graphical representations. We will start our analysis with histograms to understand the distribution of each variable and initially identify which objects are skewed or have strong outliers. Next, a matrix of scatter plots, Parallel Coordinates plot, Andrew curves and Radial visualization (Rad-Viz) plot will be analyzed.

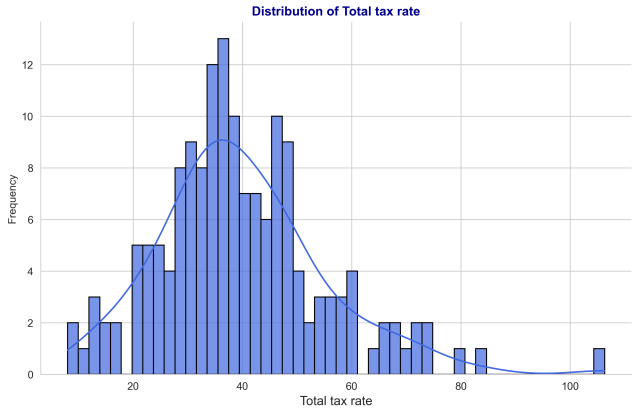

Distribution of Density(P/Km2)
Frequency
Density(P/Km2)

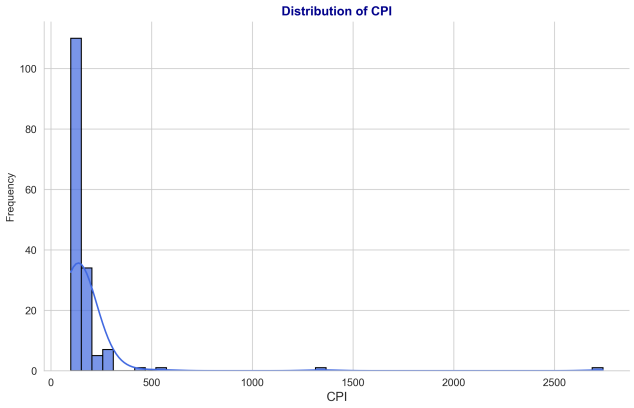

Distribution of Armed Forces
Frequency
Armed Forces
1e6

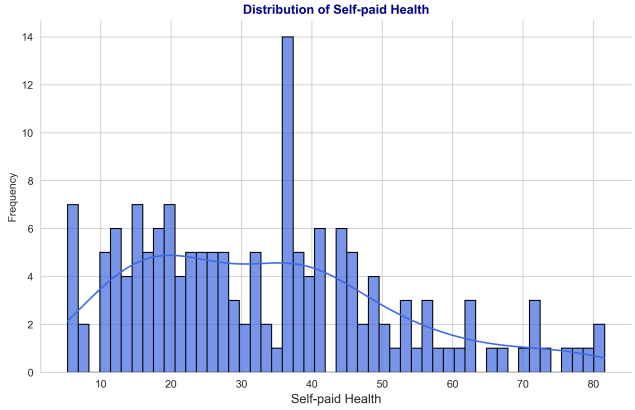

Distribution of GDP
Frequency
GDP
1e13


Distribution of Labor force participation(%)
Frequency
Labor force participation(%)


Distribution of Unemployment rate
Frequency
Unemployment rate


Distribution of Total tax rate
Frequency
Total tax rate


Distribution of CPI
Frequency
CPI


Distribution of Self-paid Health
Frequency
Self-paid Health

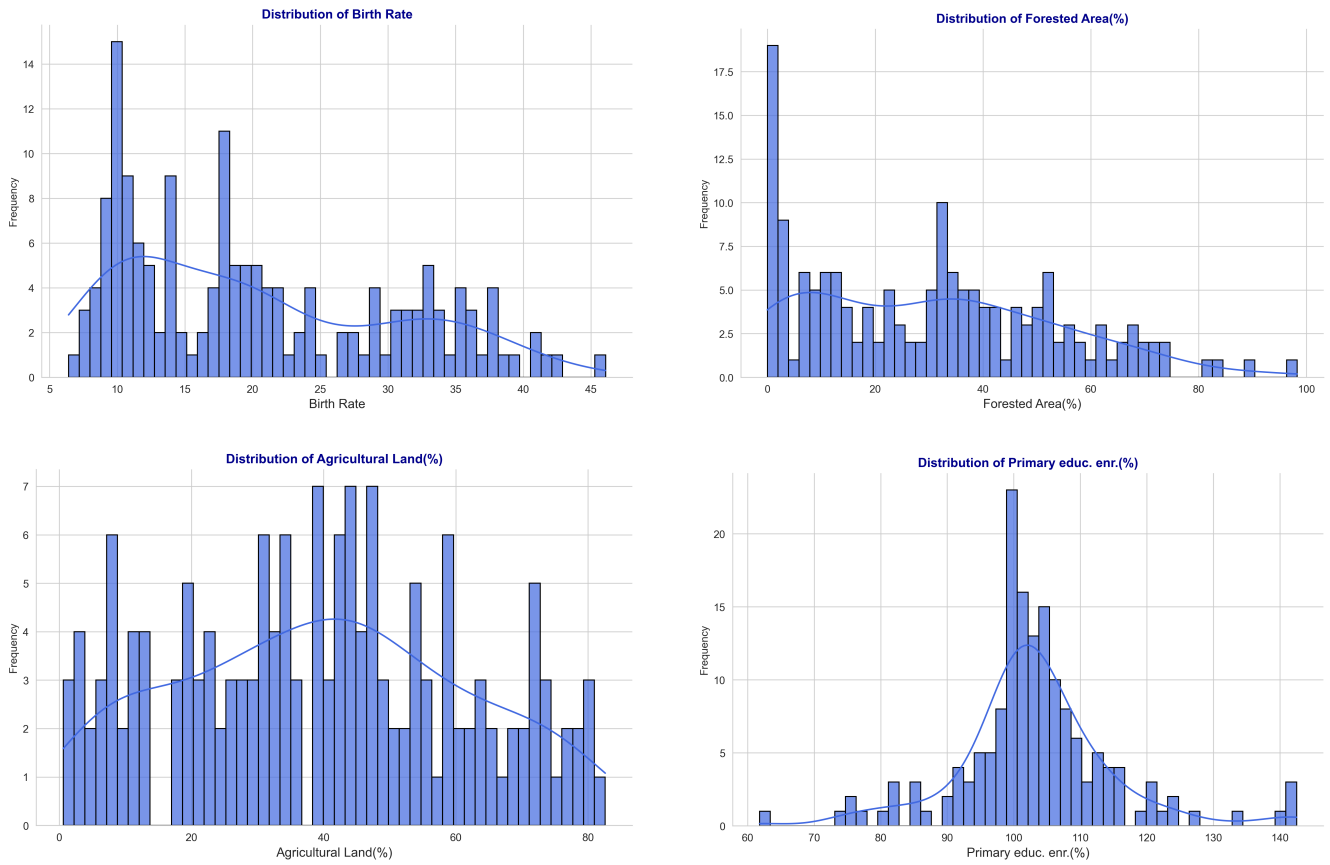


Figure 11: Distributions of variables in the original dataset

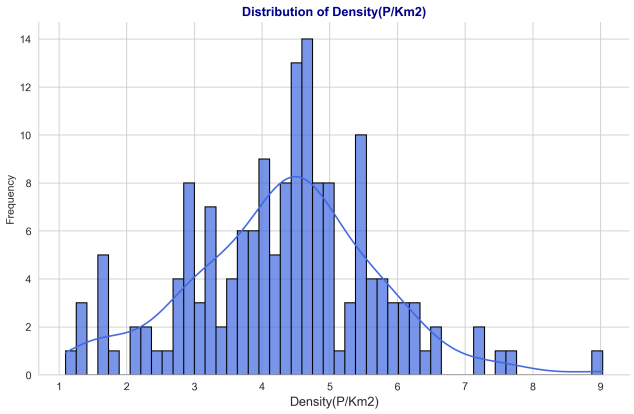
As we can see from the graphs above, some distributions are strongly influenced by outliers, making interpretation more complex. Among these variables, we find Density, Armed Forces, GDP, and CPI. While large differences in the data are normal, we can apply techniques to reduce noise and make the graphs easier to interpret. So we can consider transforming these variables using a logarithmic scale, which helps compress extreme values and makes skewed distributions more manageable. This approach reduces the impact of outliers while preserving the overall trends in the data. So we will analyze the plots of the variables in the following plot, keeping in mind that Density, Armed Forces, GDP, and CPI have to be interpret as logarithmic scale variables

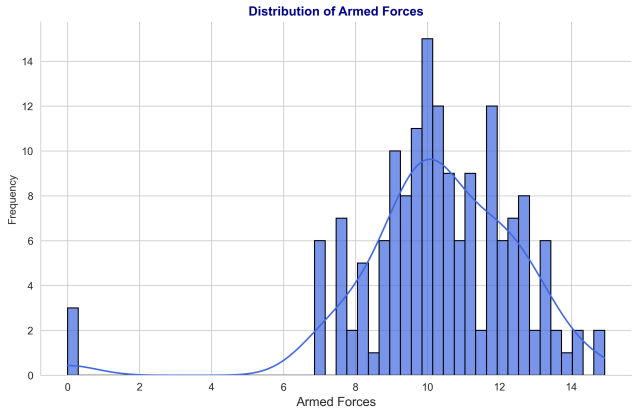
	Country	Density(P/Km2)	Armed Forces	GDP	Labor force participation(%)	Unemployment rate	Total tax rate	CPI	Self-paid Health	Birth Rate	Forested Area(%)	Agricultural Land(%)	Primary educ. enr. (%)
0	Afghanistan	4.110874	12.685411	23.673025	48.9	11.12	71.4	5.016617	78.4	32.49	2.1	58.1	104.0
1	Albania	4.663439	9.105091	23.449685	55.7	12.33	36.6	4.787908	56.9	11.78	28.1	43.1	107.0
2	Algeria	2.944439	12.666660	25.858995	41.2	11.70	66.1	5.026246	28.1	24.28	0.8	17.4	109.9
4	Angola	3.295837	11.669938	25.273298	77.5	6.89	49.1	5.571127	33.4	40.73	46.3	47.5	113.5
6	Argentina	2.890372	11.561725	26.831765	61.3	9.79	106.3	5.454252	17.6	17.02	9.8	54.3	109.7
...	...	...	...	...	...	...	...	...	...	...	...	...	...
190	Venezuela	3.496508	12.745489	26.901955	59.7	8.80	73.3	7.916177	45.8	17.88	52.7	24.5	97.2
191	Vietnam	5.752573	13.165425	26.291310	77.4	2.01	37.6	5.103032	43.5	16.75	48.1	39.3	110.6
192	Yemen	4.043051	10.596660	24.015927	38.0	12.91	26.6	5.066259	81.0	30.45	1.0	44.6	93.6
193	Zambia	3.258097	9.680406	23.861570	74.6	11.43	15.6	5.362747	27.5	36.19	65.2	32.1	98.7
194	Zimbabwe	3.663562	10.839601	23.788560	83.1	4.95	31.6	4.668239	25.8	30.68	35.5	41.9	109.9

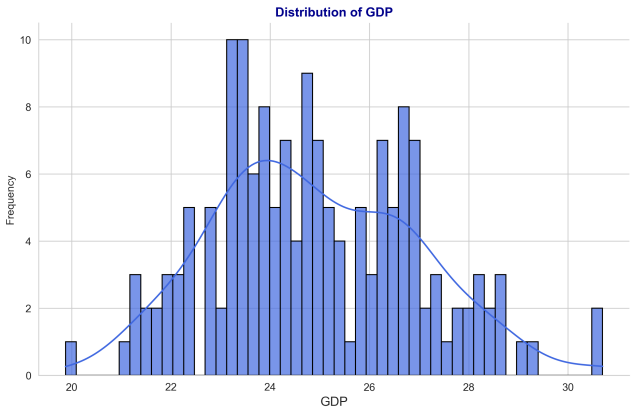
160 rows × 13 columns

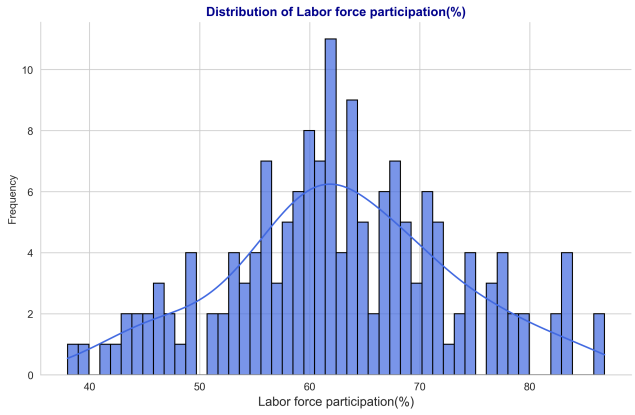
Figure 12: Dataset after logarithmic transformation of Density, Armed Forces, GDP and CPI

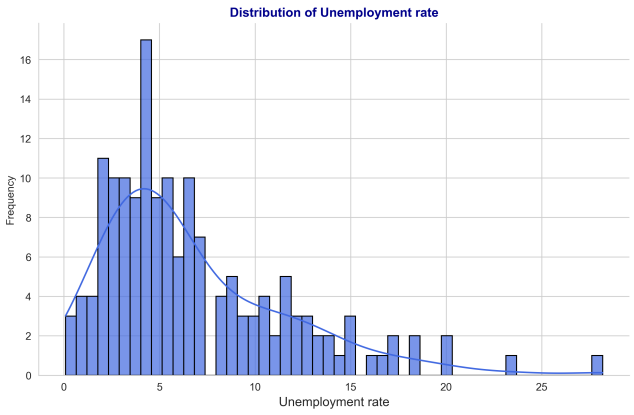
We can therefore see from figure 12 the dataset after having transformed Density, Armed Forces, GDP, and CPI to log scale variable. In the next histograms these features will be represented graphically and their distributions will be interpreted.

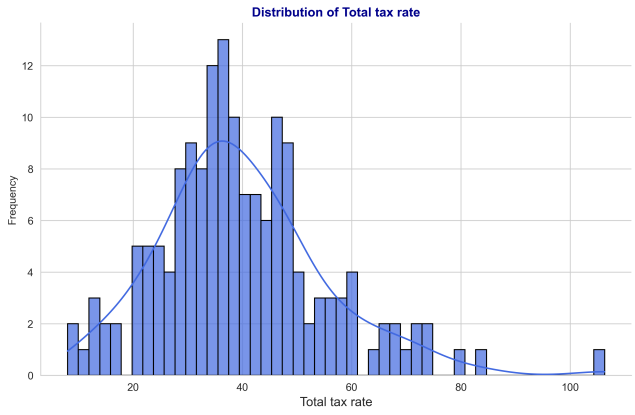

Distribution of Density(P/Km2)
Frequency
Density(P/Km2)

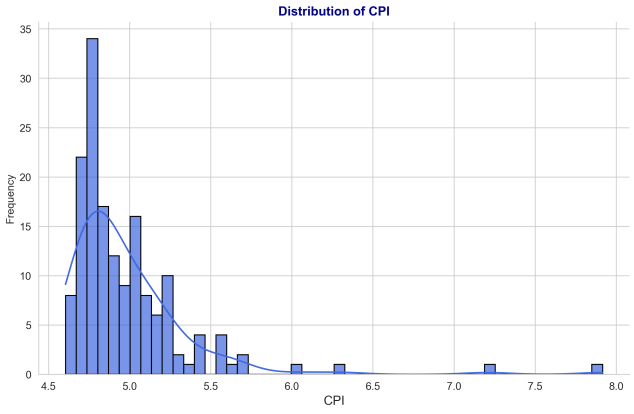

Distribution of Armed Forces
Frequency
Armed Forces

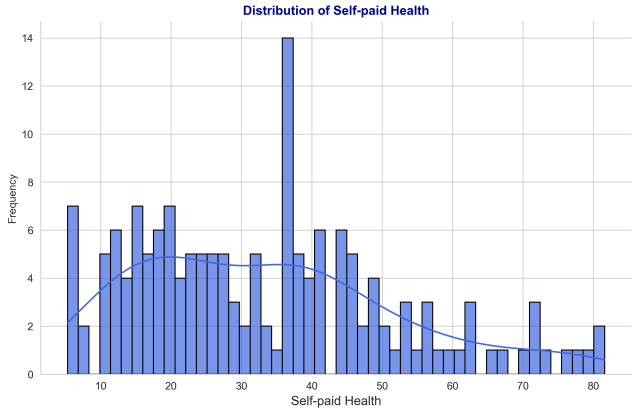

Distribution of GDP
Frequency
GDP


Distribution of Labor force participation(%)
Frequency
Labor force participation(%)


Distribution of Unemployment rate
Frequency
Unemployment rate


Distribution of Total tax rate
Frequency
Total tax rate


Distribution of CPI
Frequency
CPI


Distribution of Self-paid Health
Frequency
Self-paid Health

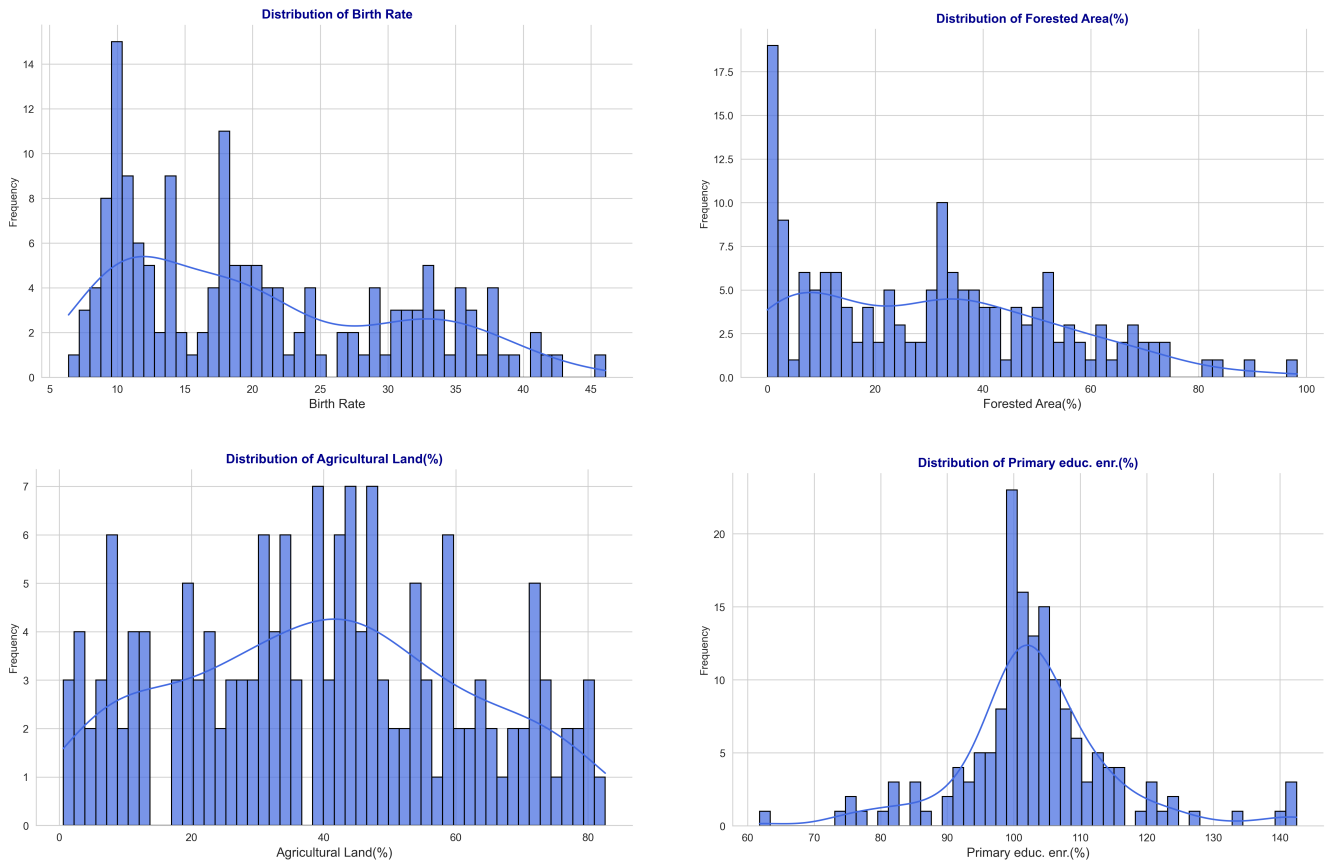


Figure 13: Distributions of variables after logarithmic transformation of Density, Armed Forces, GDP and CPI

Therefore, now the study of the graphs is much easier and more understandable. Let's move on to the analysis of each graph:

- **Density(P/km2)**: from this graph we can see how the population density varies significantly from country to country but it follows a normal distribution. It is important to remember that this variable has been transformed into a logarithmic one. Therefore the distribution may have assumed a normal distribution following the transformation. In this case we find the country with a density equal to nine is represented by Singapore with a density of 8358 people per square kilometer, followed by Bahrain and Maldives. The last place for density is occupied by Mongolia with 2 people per square kilometer and Namibia with 3 people per square kilometer.

- **Armed Forces**: the variable relating to the number of people who are part of the armed forces(in thousands) is also on a logarithmic scale. In this case we can see three anomalous observations that are represented by Libya, Iceland and Haiti which present values equal to zero, while in the first positions as armed forces we find India, China and Russia

- **GDP**: this variable has a fairly normal distribution although the presence of some outliers can be noted, including the United States and China.

- **Labor force participation(%)**: this variable has not been transformed so the values represented in the graph can be directly interpreted. The distribution is almost normal. Among the values with the highest percentage of workforce we find Qatar, Madagascar, Nepal and Rwanda. These country's economy relies heavily on subsistence agriculture, where a large portion of the population, including women and young people, engage in informal work that is still counted in labor statistics. Additionally, labor force participation does not necessarily mean formal or well-paid jobs, as informal and precarious work is also included. Many other countries with agricultural economies, such as Madagascar, Nepal, and Tanzania, show similar rates because work is often a necessity for survival rather than a structured employment opportunity.

- **Unemployment rate**: this variable, which represents the unemployment rate in percentage shows a distribution with a left tail, indicating a low unemployment rate for most of the countries. The one with the highest unemployment rate is South Africa, followed by Lesotho and Namibia while among the lowest rates we find Qatar, which previously had the highest percentage of the employed workforce.
- **Total tax rate**: the distribution of this variable appears to be normal, with an average around 40%. Among the outliers we find Argentina, where the percentage rises to 106.3% while among the countries where the tax wedge is lower we find Brunei, Georgia and Qatar.
- **CPI**: this variable has been previously transformed with a logarithmic function. From the distribution, we can see the presence of many outliers that present very high CPI values. Among these we find Venezuela, Sudan, Iran, Malawi and Suriname.
- **Self-paid health**: the distribution does not appear to be normal and shows how countries in the world have different regulations regarding the health sector. The countries where the percentage of expenditure is mostly supported by private citizens are Armenia and Yemen, while the countries with the lowest values are Botswana, Papua New Guinea, Brunei and Qatar
- **Birth rate**: the distribution in question does not show a normal trend. Countries with higher values are Niger, Chad and Mali while Japan, Spain, Italy and South Korea show much lower values
- **Forested area(%)**: regarding this variable, it can be said that the distribution is not normal and many countries have low values of areas covered by forests. In fact, many of these are countries where the percentage of territory covered by deserts is very high, such as Libya, Egypt, Qatar and Oman
- **Agricultural land(%)**: the distribution is not normal but quite constant and countries where the percentage of exploited agricultural land is higher are Uruguay and Saudi Arabia, while Singapore and Suriname have the lowest percentage
- **Primary education enrollment(%)**: this distribution follows a normal pattern with only a few outliers. Madagascar, Malawi and Nepal have very high enrollment values in primary education while countries such as Equatorial Guinea, Niger and Djibouti have low enrollment values.

2.1 Clustering and Data Normalization

Now that we have completed these first introductory analyses we can proceed with analyzing more complex graphs, including a scatter plot matrix, a Parallel Coordinates plot, Andrews curves and the Radial visualization (Rad-viz) plot. Since we are analyzing 160 states belonging to the whole world, to facilitate the visualization of the graphs we will agglomerate more countries in clusters, creating an additional categorical variable in our dataset that will indicate which cluster each country belongs to. The subdivision will be carried out using the k-means algorithm. K-means is a simple yet powerful clustering algorithm that works by grouping data points into a predefined number of clusters. It starts by randomly selecting a few points as initial cluster centers(called centroids). Then, it assigns each data point to the nearest centroid, forming clusters. Once all points are assigned, the algorithm recalculates the centroids as the average position of the points in each cluster. This process repeats iteratively, reassigning points and updating centroids, until the clusters stabilize and no significant changes occur.

To make the k-means results more stable and make subsequent visualizations more robust we will normalize the dataset by varying the values of each variable in a range between 0 and 1. In this way the variables will maintain the same trend that we analyzed in the previous paragraph in figure 11 but they will all have the same scale of variation. The formula used for the normalization of the dataset is the following:

$$X_i' = \frac{X_i - X_{\min}}{X_{\max} - X_{\min}}$$

- X_i is the original value of the variable for observation i.
- X_i' is the normalized value of the variable for observation i.
- $X_{\min}$ is the minimum value of the variable in the dataset.
- $X_{\max}$ is the maximum value of the variable in the dataset.

We then proceed with the cluster analysis where we will divide our countries into 3 main clusters. By noting the composition of these clusters and analyzing the normalized average GDP of each cluster we can divide our observations into:

- rich: the normalized mean of GDP is 0.513
- medium_rich: the normalized mean of GDP is 0.419
- poor: the normalized mean of GDP is 0.246

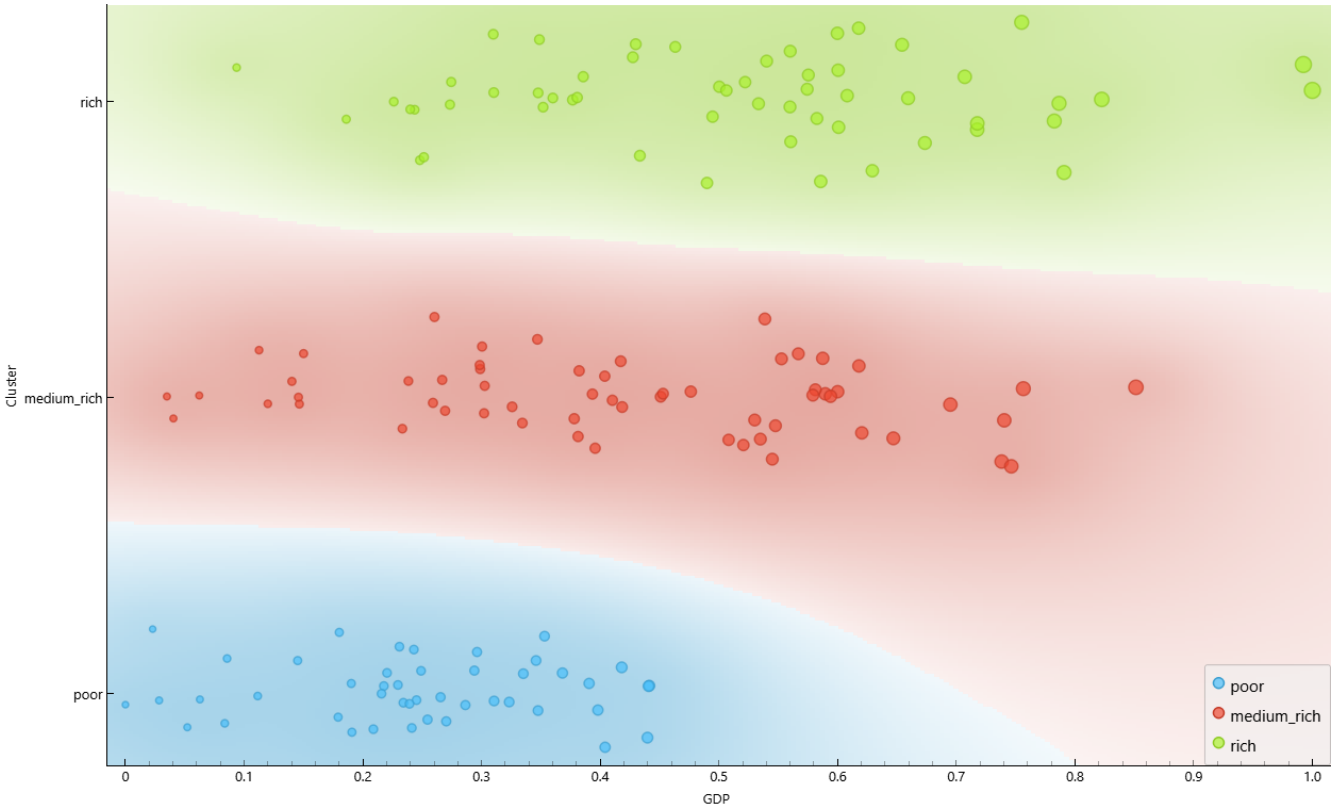


Figure 14: Scatter plot of clusters related to GDP

From the graph in figure 14 we can see the subdivision of the observations into three distinct clusters. The size of each dot depends on the GDP of the country and obviously we can see that some countries classified as medium rich could easily fall into the cluster of poor countries just as some could fall into the cluster of rich countries.

In the next paragraphs we will initially analyze the scatter plots and then we will present the Parallel Coordinates plots, the Andrews curves and the Radial Visualization plot

2.2 Scatter plot matrices

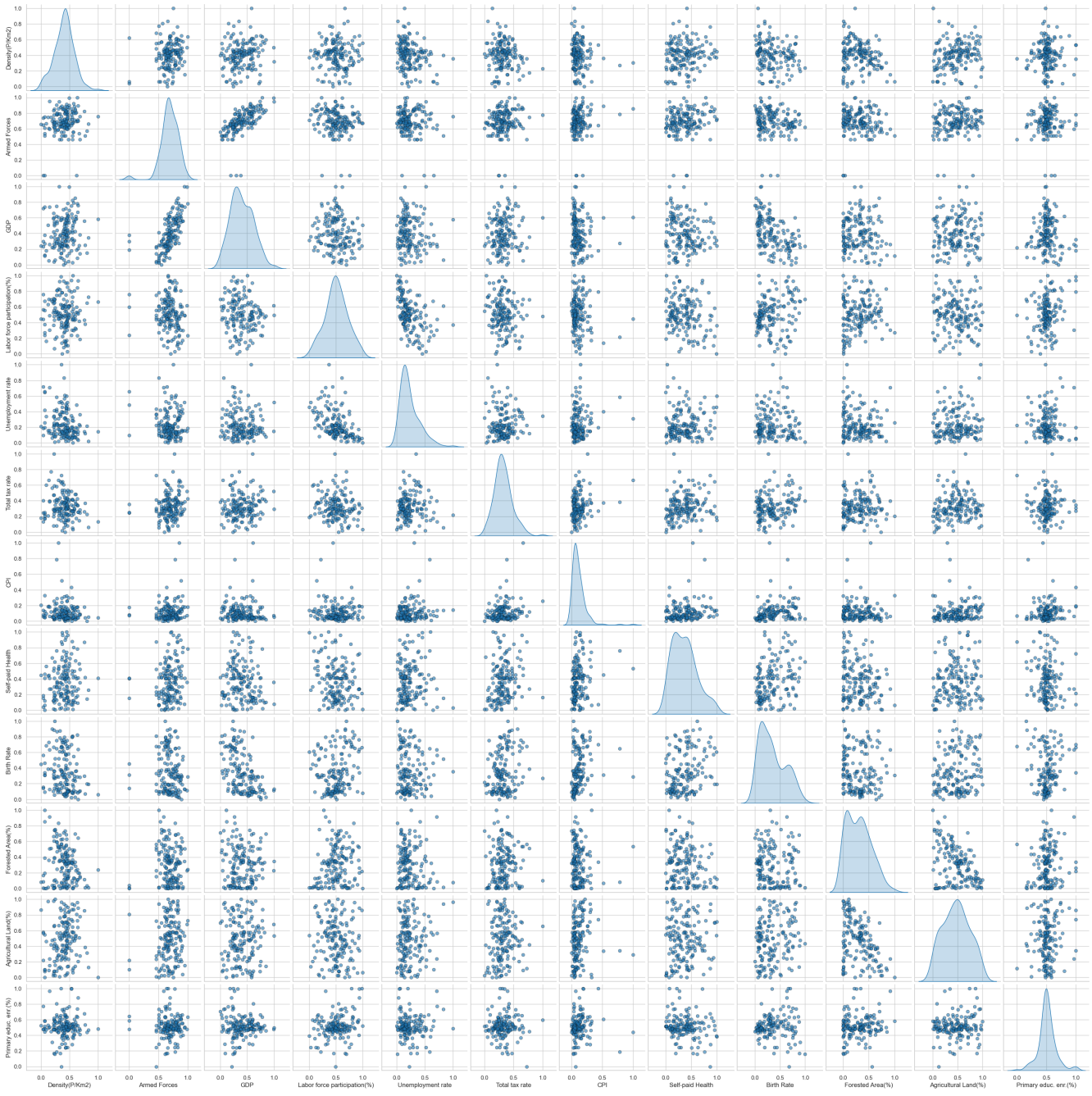


Figure 15: Scatter plot matrix of every variables in the dataset

From figure 15 we can see how each variable relates to the corresponding variables. The graph is a bit confusing due to the high number of variables analyzed but it can be noted in general that there are no strong correlations between one variable and another. Only the GDP variable appears to be slightly correlated with the variable 'Armed forces' as the correlation is 63%. Similarly, variables such as 'Labor force participation' and 'Unemployment rate' are negatively correlated at 47% and 'Agricultural Land' and 'Forested Area' are negatively correlated at 42%. To facilitate the understanding of this scatter plot matrix we will divide the dataset into two macro categories, namely economic and non-economic variables. The first group includes 'Armed Forces', 'GDP', 'Labor force participation', 'Unemployment rate', 'Total tax rate' and 'CPI' while the second category includes 'Density(P/km2)', 'Self-paid Health', 'Birth rate', 'Forested Area', 'Agricultural Land' and 'Primary education enrollment'.

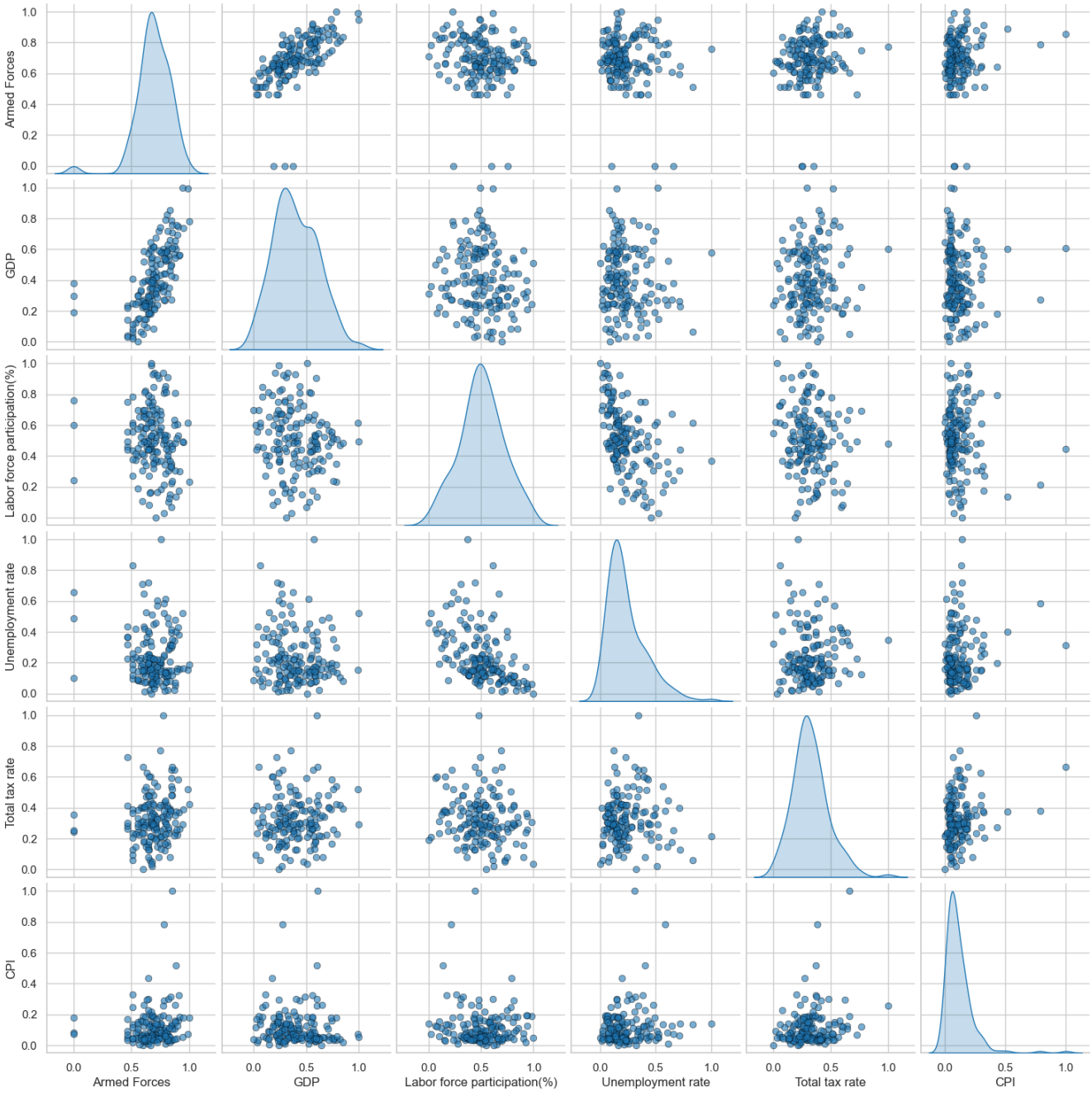


Figure 16: Scatter plot matrix of economic variables

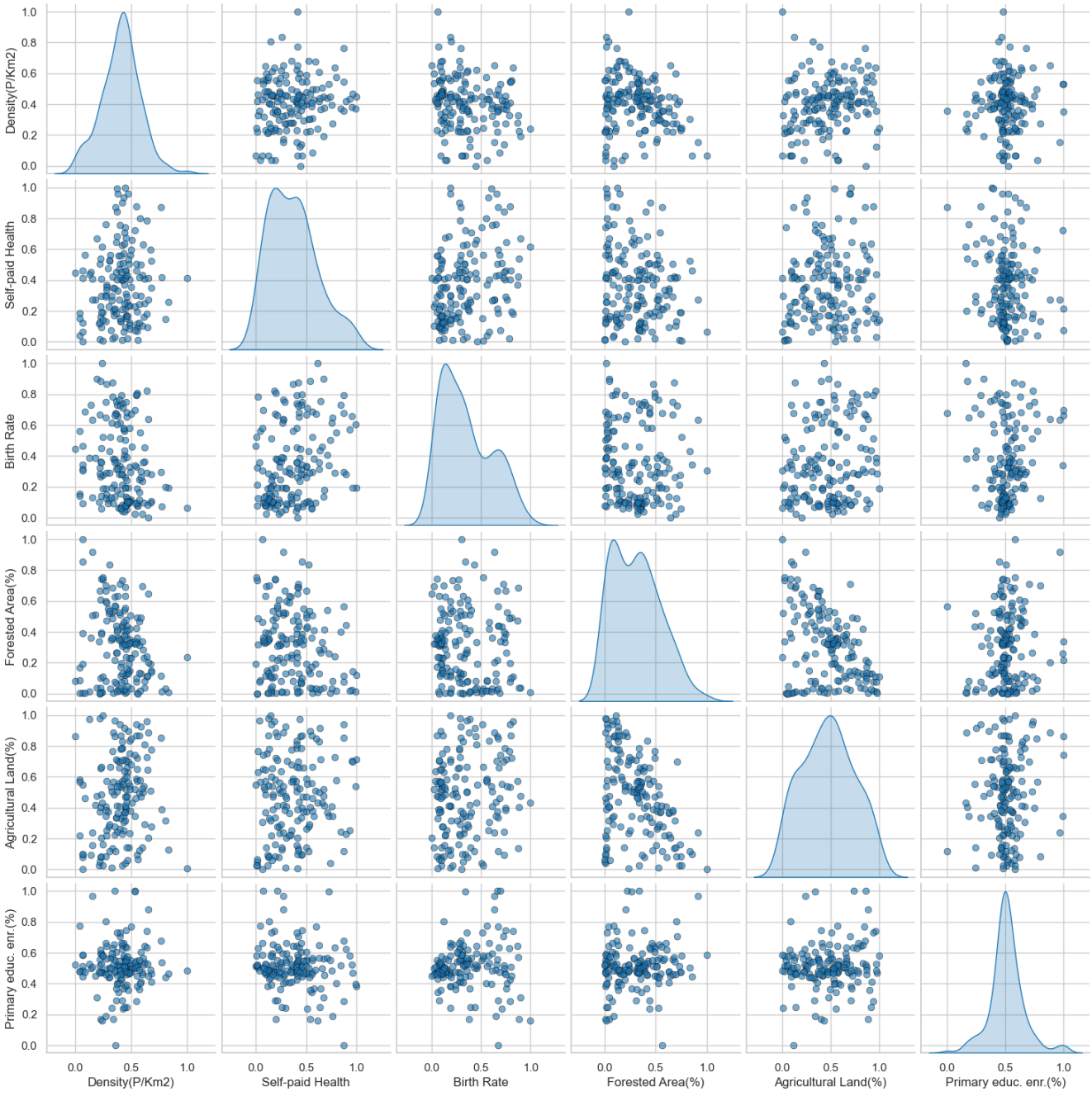


Figure 17: Scatter plot matrix of non-economic variables

2.3 Parallel Coordinates plot

From figure 18 relating to the Parallel Coordinates plot we can observe that if we considered all the countries in the same graph, the latter would be difficult to understand and noisy. Therefore, considering the averages of each cluster we can note the general trend of the various countries divided into groups. In particular, we can observe that the richest countries tend to be, on average, the most densely populated and those with the highest levels of investment in the military sector.

High population density in wealthy nations can be attributed to advanced urbanization, well-developed infrastructure, and economic opportunities that attract both internal migration and immigration. These countries often have large metropolitan areas that serve as global economic hubs, further increasing population concentration.

Additionally, greater military investments in rich countries can be linked to their geopolitical influence, national security strategies, and technological advancements in defense. With larger economies, these nations allocate significant resources to maintaining powerful armed forces, developing innovative military technology, and sustaining global defense commitments. This correlation between wealth, population density, and military expenditure highlights how economic strength often translates into both strategic defense priorities and the capacity to support a large and concentrated population. This is confirmed by the GDP which once again underlines the division of the clusters based on the wealth of each country.

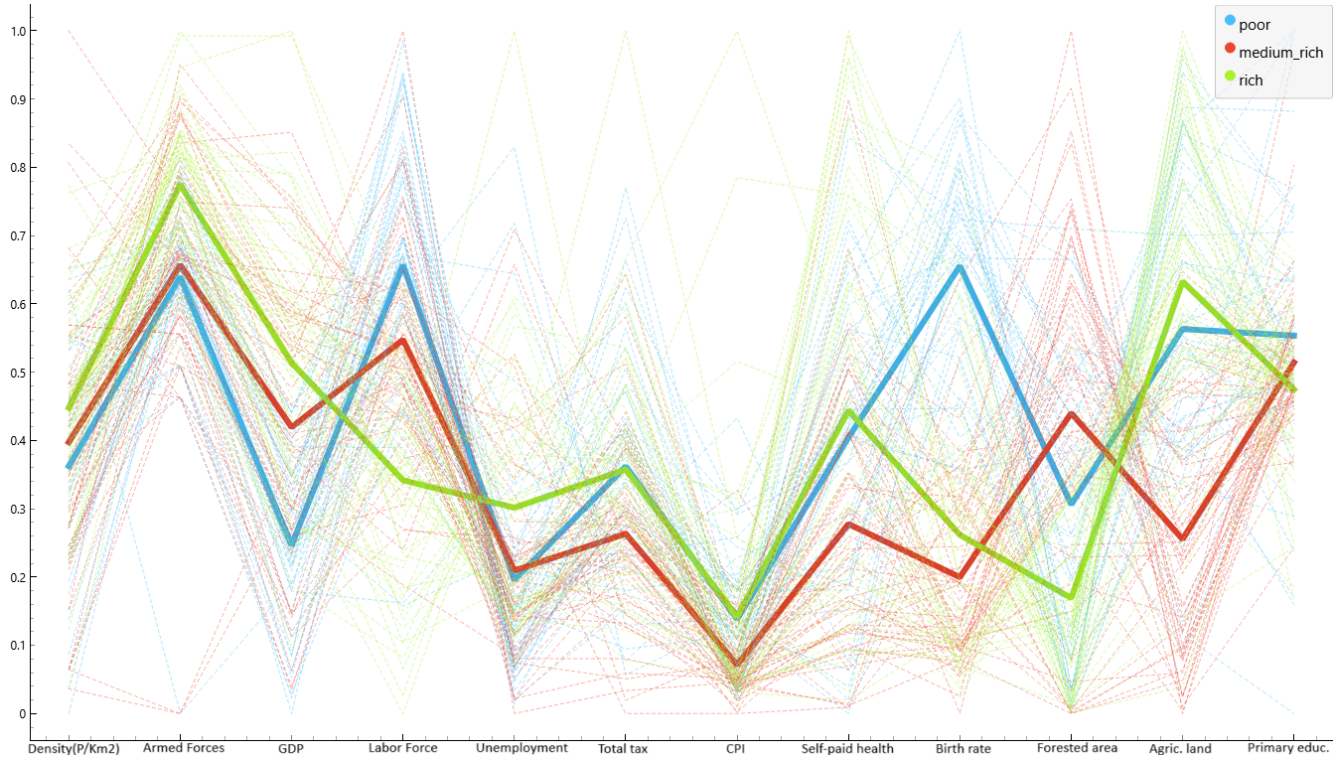


Figure 18: Parallel Coordinates plot

Continuing we have the labor force variable in which we find the poorest countries at the top, and this underlines the same concept expressed previously when we analyzed the individual distributions of each variable in figure 13. In poorer countries, a larger share of the population is employed compared to richer nations due to several structural and economic factors. One of the main reasons is the lack of welfare systems, which means that most people cannot afford to be unemployed. Unlike in wealthier countries, where social safety systems like unemployment benefits and pensions allow some individuals to stay out of the workforce, in poorer economies, almost everyone needs to work to survive. Additionally, a significant portion of employment in these countries is informal, meaning that many people work without contracts, stable wages, or labor protections. This creates the illusion of a high employment rate, even though the jobs are often low-paying and unstable. Another key factor is the prevalence of subsistence agriculture, where many people work on family farms, not necessarily earning a salary but contributing to household survival. In contrast, in richer countries, technological advancements and higher productivity allow fewer workers to generate more wealth, reducing the need for a large workforce. Moreover, education plays a crucial responsibility in developed nations, people spend more years in school before entering the job market, in poorer countries, many

start working at a young age to support their families. Women's participation in the informal sector is also higher in less developed economies, often through domestic work, street vending, or small-scale farming, further increasing the apparent labor force participation rate. Ultimately, while employment rates may be higher in poorer countries, the quality of jobs, wages, and working conditions are significantly lower compared to wealthier nations. The unemployment rate also confirms this problem, indicating that poorer countries have fewer unemployed than rich countries, however as previously stated it is not always synonymous with better working conditions and higher earnings.

Let's continue with the variable relating to taxes or rather the overall burden of taxation on businesses, expressed as a percentage of trade profits. In this case on average it seems that poor and rich countries support the same tax wedge, unlike the average rich countries that differ from these by not positioning themselves in the middle between the two clusters.

Regarding the CPI, we know that an increase in the CPI indicates rising prices, signaling inflationary pressures in the economy while a value that is too low (or negative) may suggest deflation. Therefore, regarding our clusters we can say that rich and poor countries have on average the same value. If both rich and poor countries have the same Consumer Price Index (CPI), it does not necessarily mean that their economies are similar, but it can indicate several underlying factors. When a wealthy and a developing nation share the same CPI value, it could suggest that both are experiencing similar inflation rates, but the reasons behind it may differ. In a richer country, inflation might result from strong consumer demand, rising wages, or supply chain disruptions, while in a poorer nation, it could be driven by currency depreciation, higher import costs, or structural inefficiencies. Additionally, even if CPI values are identical, the impact on the population varies significantly. In wealthier countries, people typically have higher wages and better purchasing power, so moderate inflation may not severely affect their standard of living. In contrast, in developing economies, where incomes are lower and a large portion of household spending goes to necessities like food and housing, even small price increases can have a much harsher impact. Therefore, while a matching CPI value might indicate a similar inflation rate, it does not imply economic equality or comparable living conditions between rich and poor nations.

Proceeding we find the variable relating to the percentage of health care expenditure that is paid directly by citizens, without coverage by the state or insurance. It is possible to see how the value is higher for rich and poor countries and lower for medium-rich countries. This pattern may be attributed to different healthcare system structures. In wealthier countries, while many have strong public or private insurance systems, certain expenses, such as out-of-pocket payments for specialized treatments, private healthcare services, or uncovered procedures, can still be significant. Conversely, in poorer countries, limited government funding and underdeveloped insurance systems force individuals to bear a large share of healthcare costs directly. Middle-income countries, on the other hand, often have expanding healthcare systems that provide more coverage and protection against high out-of-pocket expenses, leading to a lower direct financial burden on citizens.

Moving on to the variable concerning the number of births per 1,000 inhabitants per year, we can observe that birth rates are significantly higher in poorer countries compared to wealthier ones. This trend is largely influenced by economic, social, and cultural factors. In less developed nations, limited access to contraception, lower levels of education and the need for larger families to provide economic support contribute to higher fertility rates. Additionally, higher infant mortality rates often lead families to have more children as a form of security. In contrast, in richer countries, greater access to education, widespread availability of contraception, higher living costs, and lifestyle changes have led to a decline in birth rates, with many families choosing to have fewer children. This demographic divide highlights how economic development and social policies influence population growth across different regions of the world.

Regarding the variables related to the environmental aspect, we observe that rich and presumably more industrialized countries tend to have a lower percentage of land covered by forests compared to poor and middle-income countries. This is largely due to extensive urbanization, industrial expansion, and agricultural development, which have historically led to deforestation in wealthier nations. However, what stands out is that middle-income countries often have a higher percentage of forested areas than poorer nations. This could be attributed to several factors, such as increased environmental awareness, government policies promoting reforestation, and better land management practices. In contrast, poorer countries, despite having vast natural landscapes, often face significant deforestation due to economic pressures, illegal logging, and the need for agricultural expansion to support growing populations. This suggests that while industrialized nations have already undergone major deforestation, and middle-income countries are increasingly prioritizing conservation efforts, poorer nations may still be struggling to balance economic development with environmental sustainability.

When looking at the variable related to the percentage of a country's territory used for agriculture, we find that rich countries tend to have a higher percentage of agricultural land compared to both poor and middle-income

nations. This may seem counterintuitive, but it can be explained by the highly developed and efficient agricultural sectors in wealthier nations. Advanced technology, large-scale farming, and extensive infrastructure allow these countries to maximize land use for agricultural production, often for both domestic consumption and export. In contrast, poorer countries, despite having vast rural areas, may lack the necessary resources, technology, and infrastructure to cultivate large portions of land efficiently. Additionally, land degradation, deforestation, and conflicts over land ownership can further limit agricultural expansion. Middle-income countries, on the other hand, may be experiencing a transition where urbanization and industrialization reduce the share of land dedicated to agriculture, while improvements in efficiency allow for more productive use of smaller areas. This trend highlights how economic development influences land use, with wealthier nations able to sustain large-scale agricultural activities while less developed countries may struggle to fully exploit their land potential.

Finally, we conclude with the variable related to education, specifically the gross enrollment ratio for primary education, which represents the percentage of children enrolled in primary school compared to the population of primary school age. Interestingly, we observe that poorer countries tend to have slightly higher enrollment rates than richer nations. This may be due to efforts by governments and international organizations to promote universal primary education, often making enrollment almost mandatory and heavily subsidized in developing countries. Additionally, in many wealthier nations, factors such as alternative education paths, homeschooling, or slightly different age classifications for school entry may result in a marginally lower enrollment ratio. However, while the enrollment rate might be high in poorer countries, challenges such as school dropout rates, teacher inadequacy, and lower education quality can still obstruct overall educational outcomes. This suggests that while access to primary education has improved globally, the quality and effectiveness of education systems remain key differentiating factors between rich and poor countries.

2.4 Radial visualization(Rad-Viz) plot

Observing the plot in figure 19, we can clearly see that the data reveals a concentric arrangement with some overlap between the groups, yet each maintains well-defined characteristics. Poor countries, shown in blue, exhibit greater dispersion, indicating a wider variability in their socioeconomic indicators. This suggests that within this category, there are significant differences between nations but they all share common factors such as high birth rates.

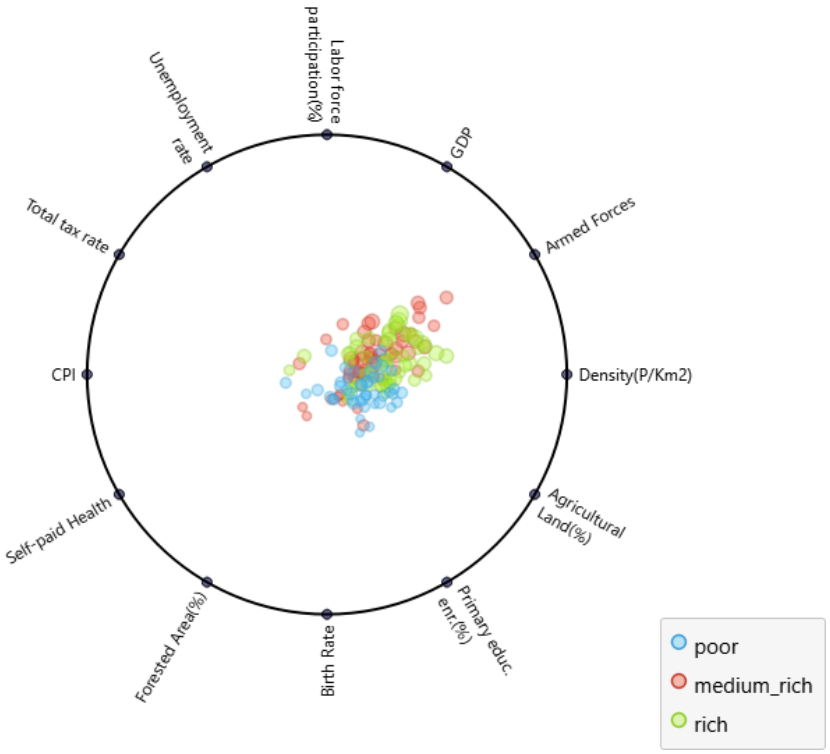


Figure 19: Rad-Viz plot

Medium-rich countries, represented in red, occupy a more central position in the chart, acting as a bridge

between the two extremes. Their distribution suggests a higher degree of homogeneity compared to poor countries, yet still some variability, indicating that these nations are in a phase of economic transition or development. Finally, rich countries, shown in green, tend to cluster in a more compact and concentrated area, suggesting greater stability across the indicators analyzed. This uniformity implies that wealthier nations share more similar characteristics, generally exhibiting better levels of GDP, armed forces, education, and other key factors.

In summary, the chart highlights not only the distinction between these three groups but also the degree of dispersion and homogeneity within each. It clearly shows that wealthier countries tend to be more alike, while poorer nations display a much broader range of variations in their economic and social indicators.

2.5 Andrew curves

This graph presents a series of overlapping line plots that illustrate the behavior of rich, poor, and medium-rich countries, represented by dark green, light green, and yellow, respectively. The lines exhibit a common oscillatory pattern, suggesting that all groups follow similar trends, though with variations in amplitude and density. Poor countries, shown in light green, appear to have a slightly wider spread, indicating greater variability within this group. Medium-rich countries, in yellow, display an intermediate behavior, bridging the gap between the two extremes. Meanwhile, rich countries, depicted in dark green, tend to cluster more tightly, implying greater stability and consistency in their values.

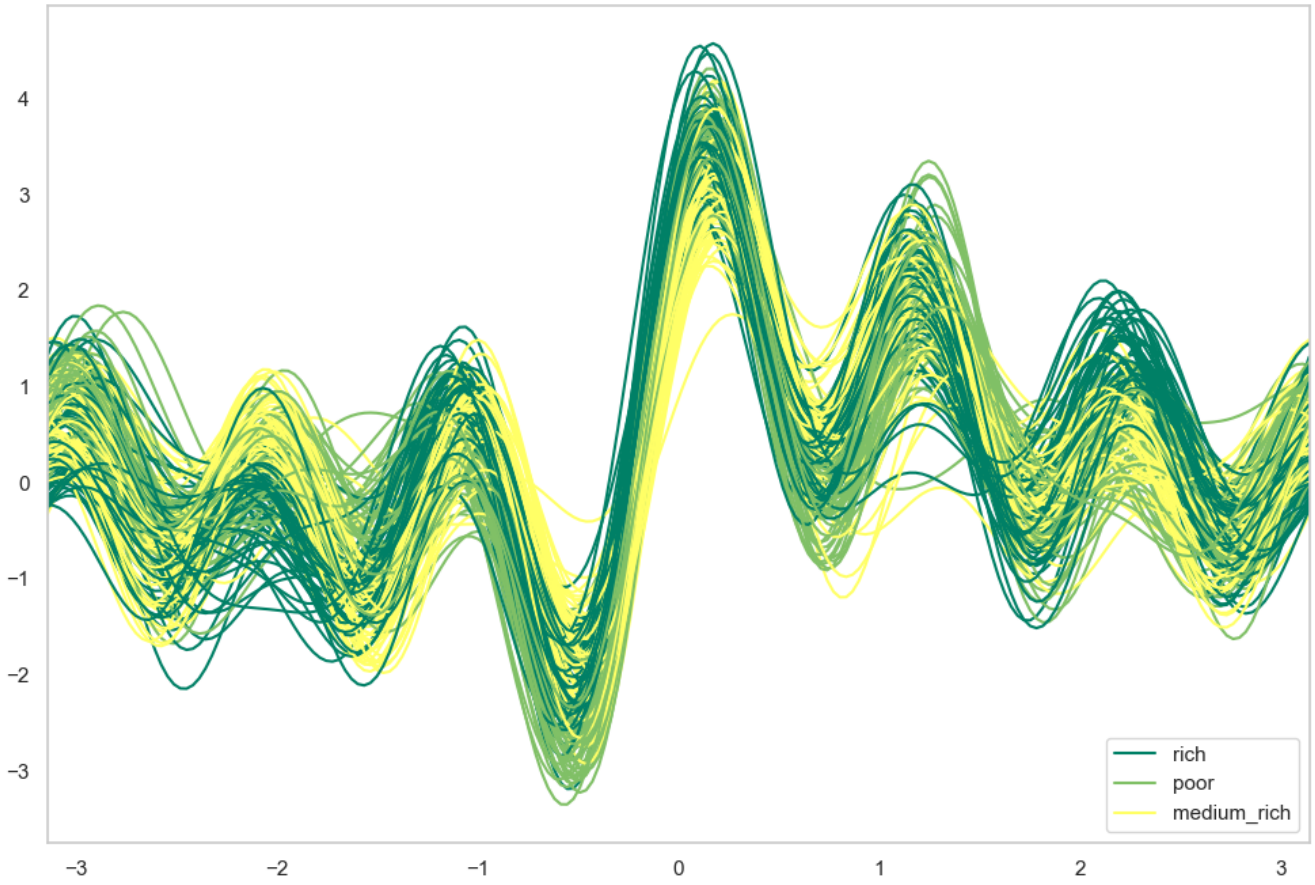


Figure 20: Andrew curves

Despite the differences in dispersion, the shared peaks and troughs across all groups suggest that they respond to similar underlying factors, although to varying degrees. This structure highlights how different economic or social conditions influence each category in a comparable yet distinct manner. The more compact distribution of rich countries indicates a more uniform economic or social environment, whereas the wider spread of poor nations suggests a broader range of disparities. By visualizing these variations, the graph effectively conveys the complex differences and similarities among the three groups, providing insight into the dynamics that shape their respective trajectories.

2.6 Conclusions on Data visualization

All the methods used to analyze the data provided great information about the dataset and helped deepen our analysis. The graphs contributed differently to the study but the Parallel Coordinates plot in figure 18 was the most useful to understand how on average each country cluster differs from the others. This allowed us to identify trends and patterns among the variables, compare the different categories and analyze the distribution of the data in multiple dimensions.

3 Principal Component Analyses

We will now perform the Principal Component Analysis on our original dataset, that is, the dataset composed of 26 variables and summarized in figure 6 in the first chapter. Since PCA is a dimensionality reduction technique that is based on the explained variance, it makes sense to apply it on the dataset that includes all the initial variables rather than on the dataset previously used for data visualization. This will allow us to compare the results obtained automatically through PCA with those obtained by personally choosing the variables based on the study of the correlation and economic significance. The dataset used will therefore have 26 columns and 120 observations, each of which represents a state. In this case, the number of instances drops from 160 to 120 because we had to eliminate the missing values before performing the PCA.

3.1 The choice of the number of components

Now we need to decide how many components to use in order to best summarize our dataset. There are mainly three methods that could be used:

1. A number of components is considered such that they take into account a sufficiently high percentage of the total variance. As the number of variables increases, the total variance increases and therefore it may be reasonable to consider a smaller percentage of explained variance. It can be requested that the extracted components take into account on average at least 95% of the variance of each of the p starting variables, which implies setting the minimum threshold equal to the following expression: $0.95^p * 100$, which is equal to 77.38% if p = 5; to 59.87% if p = 10. In our case, where we are studying 26 variables, we would like the principal components to take into account on average 26.35% of the variance of each of the 26 variables.

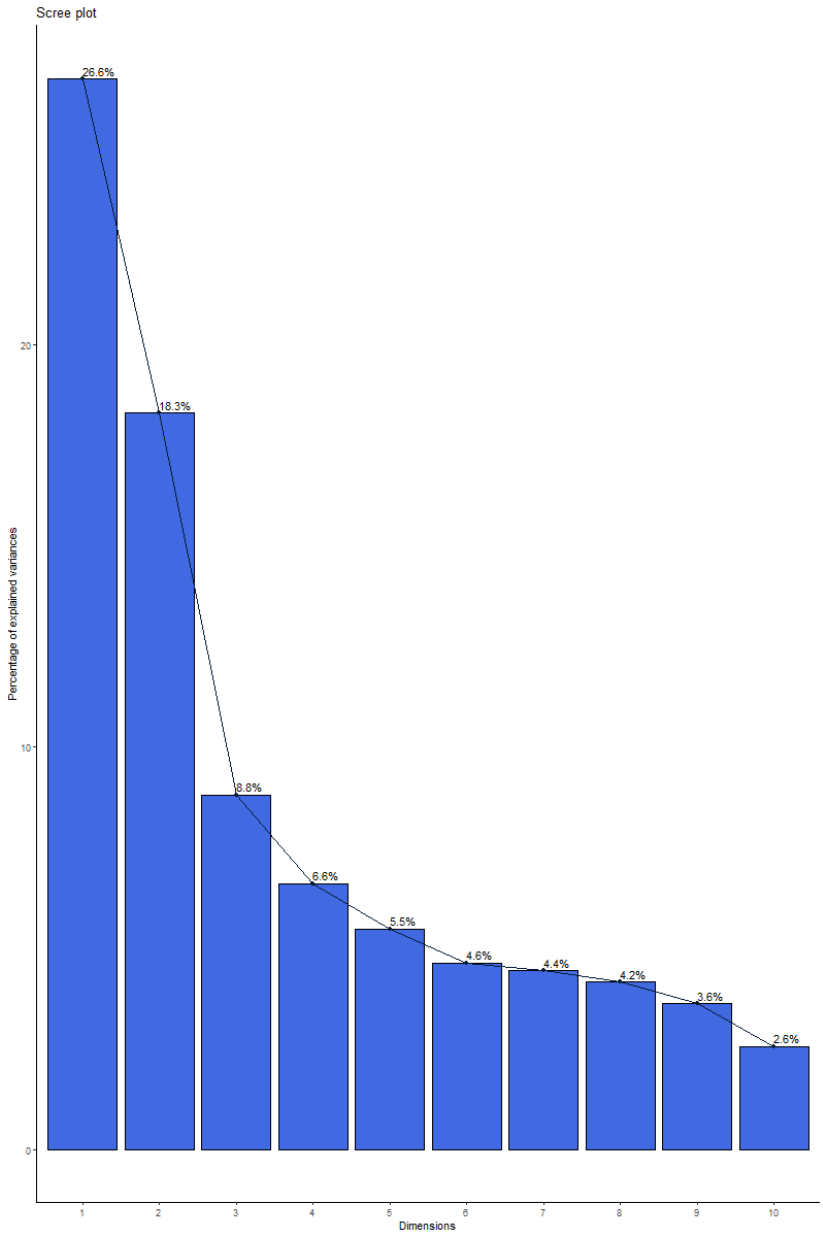
2. All components whose eigenvalue is greater than 1 are preserved. The ratio of this method derives from the fact that the eigenvalue of a component is equal to its variance and that by operating on standardized variables these have unit variance. Therefore, we decide to maintain a component (which is a linear combination of the p variables) only if it explains a greater share of the total variance than the variance explained by a single variable.

3. We can construct a scree plot of the eigenvalues as a function of 'v' components (v = 1, 2, ..., p). A scree plot is a simple line segment plot that shows the eigenvalues for each individual component. It shows the eigenvalues on the y-axis and the number of factors on the x-axis. If k components are important and the remaining (p - k) insignificant, between k and k + 1 there is a sharp variation in the slope (an "elbow"), which signals that k is the appropriate number of components to conserve. Basically, the scree plot criterion looks for the "elbow" in the curve and selects all components just before the line flattens out.

Generally it is the joint use of the three criteria that allows identifying the number k of components to use. Starting with the first method, since we have 26 variables to consider, the number of components must be such as to explain at least 26.35% of the variance of the variables. According to this method, the best number of components to use is two, with a cumulative variance of 45%. Of course the power of the PCA will be limited since the remaining part of the variance, equal to 55%, is not explained but considering the high number of variables taken into account we can accept this limitation.

Considering the second method explained before, we should consider only the components whose eigenvalue is greater than 1, since they explain more variance than a single original variable. So, by looking at table 21(b) just the first eight components must be considered, but the PCA model will be less flexible and interpretable. Also considering the scree plot in figure 21(a) the first three components should be considered since we observe a change in the slope at k equal 3. Even if in this case the change in the slope is not so marked, it can be noted how from the third component the amount of variance explained tends to decrease significantly compared to the first two

components, therefore we will take into consideration a number of components equal to 2 and subsequently we will deepen our studies with a three-dimensional graph that considers 3 components.



(a) Scree plot

Eigenvelues	Variance(%)	Cumulative variance(%)
6.913	26.590	26.590
4.756	18.290	44.880
2.288	8.799	53.680
1.716	6.600	60.280
1.422	5.469	65.748
1.203	4.628	70.376
1.154	4.437	74.813
1.080	4.155	78.968
0.939	3.613	82.582
0.663	2.551	85.133
0.623	2.395	87.527
0.529	2.033	89.560
0.478	1.840	91.400
0.390	1.499	92.899
0.350	1.346	94.246
0.345	1.327	95.573
0.294	1.131	96.704
0.247	0.950	97.654
0.195	0.749	98.403
0.180	0.692	99.095
0.085	0.326	99.421
0.072	0.277	99.698
0.035	0.136	99.834
0.026	0.101	99.935
0.010	0.039	99.974
0.007	0.026	100.000

(b) PCA summarization

Figure 21: How to choose the best number of components for the PCA ?

3.2 Performing PCA with two components

We can finally start with the principal component analysis, that will consider at first two principal components and will help us to summarize the information available. In the component table, shown in figure 22, we can find the correlation coefficients between each variable and the two main components. The sign of these coefficients indicates the type of linear relationship, direct or inverse, between the component and the variable to which they refer, while the numerical value, in module, indicates the entity of the link. As we can see from the table, the variables 'Urban population', 'Population', 'Armed forces', 'GDP' are negatively correlated with the second component and slightly negatively correlated with the first. On the contrary, we note that the variables 'Birth Rate', 'Fertility Rate', 'Infant mortality' and 'Mortality mortality ratio' are positively correlated with the first component. This can also be seen graphically in figure 23.

In figure 22 we also find the Communality column, which highlights how well the starting variables are explained by the components of the model. As we can see, some columns are explained very well, such as 'Urban population', 'Population', 'Armed forces', 'GDP', 'Birth Rate', 'Fertility Rate', 'Life expectancy', 'Infant mortality' and 'Co2-Emission'. The other variables present very low levels of communality, such as 'Density (P/km2)', 'Labor force participation (%)', 'Unemployment rate', 'Total tax rate', 'CPI', 'Self-paid Health', 'Forested Area (%)', 'Agricultural Land (%)', 'Gasoline Price' and 'Primary educ. enr. (%)'. As for these variables it is possible that there is another component that is able to group them and explain the variance better within the bi-plot. For this reason, in the next paragraph we will consider three components to evaluate whether the variability of these variables can be explained in a better way.

	Component 1	Component 2	Communality
Density(P/Km2)	-0.197	0.023	0.039
Urban population	-0.239	-0.936	0.933
Population	-0.160	-0.894	0.825
Armed Forces	-0.238	-0.889	0.847
GDP	-0.349	-0.770	0.715
Minimum wage	-0.640	0.155	0.434
Labor force participation(%)	0.259	-0.007	0.067
Unemployment rate	-0.072	0.009	0.005
Tax revenue(%)	-0.462	0.403	0.376
Total tax rate	0.102	-0.251	0.073
CPI	0.282	-0.126	0.095
CPI Change(%)	0.245	-0.165	0.087
Doctors/1000	-0.787	0.194	0.657
Self-paid Health	0.343	-0.160	0.143
Birth Rate	0.931	-0.088	0.875
Fertility Rate	0.896	-0.091	0.811
Life expectancy	-0.934	0.145	0.893
Infant mortality	0.919	-0.148	0.866
Maternal mortality ratio	0.842	-0.125	0.725
Co2-Emissions	-0.310	-0.889	0.886
Land Area(Km2)	-0.230	-0.633	0.454
Forested Area(%)	0.024	0.028	0.001
Agricultural Land(%)	0.025	-0.085	0.008
Gasoline Price	-0.279	0.247	0.139
Primary educ. enr.(%)	0.093	-0.040	0.010
Tertiary educ. enr.(%)	-0.835	0.063	0.701

Figure 22: First two components and communality

The graph in figure 23 represents the loadings of the Principal Component Analysis. These loadings help us interpret how the original variables contribute to the variation captured by the two components. In general, three distinct groups of loadings can be noted. The first is formed by the variables Birth Rate, Fertility Rate, Infant Mortality and Maternal Mortality Ratio which are positively correlated to the first component. On the contrary, the variables Life Expectancy, Doctors/1000, Tertiary Education Enrollment and Minimum Wage are negatively correlated with the first component and are inversely related to the first group of variables. Therefore, the first component is able to capture the difference between developed and non-developed countries, having on the right side of the graph the indicators related to population growth and mortality during childbirth of the mother and child. Therefore, these values, if very high in a country, could indicate a poor health system, in which there is a high birth rate but at the same time a high mortality rate and this is often associated with third world countries

that are unable to properly manage the health system.

On the contrary, on the left side of the graph we find the indicators that are associated with better health care and better education. It can be noted that higher values of these indicators, such as Life expectancy or Tertiary Education Enrollment are associated with highly developed and modern countries. This division suggests that countries scoring high on Component 1 may struggle with underdeveloped public health systems, whereas those with negative Component 1 scores exhibit characteristics of modern, industrialized economies.

Finally, we find a last group of indicators that is not correlated with the first component but rather with the second. Among these we find Land Area, GDP, Armed Forces, Population, Urban population and Co2-Emissions. These indicators, purely economic in nature, indicate rich countries with large armed forces and high Co2 emissions while more policy and welfare-oriented indicators might lie in the opposite direction. Therefore, it can be said that Component 1 mainly contrasts health and education development against high birth/mortality rates and population-driven metrics. Component 2 adds another layer regarding fiscal/social policies, density and military/economic magnitude.

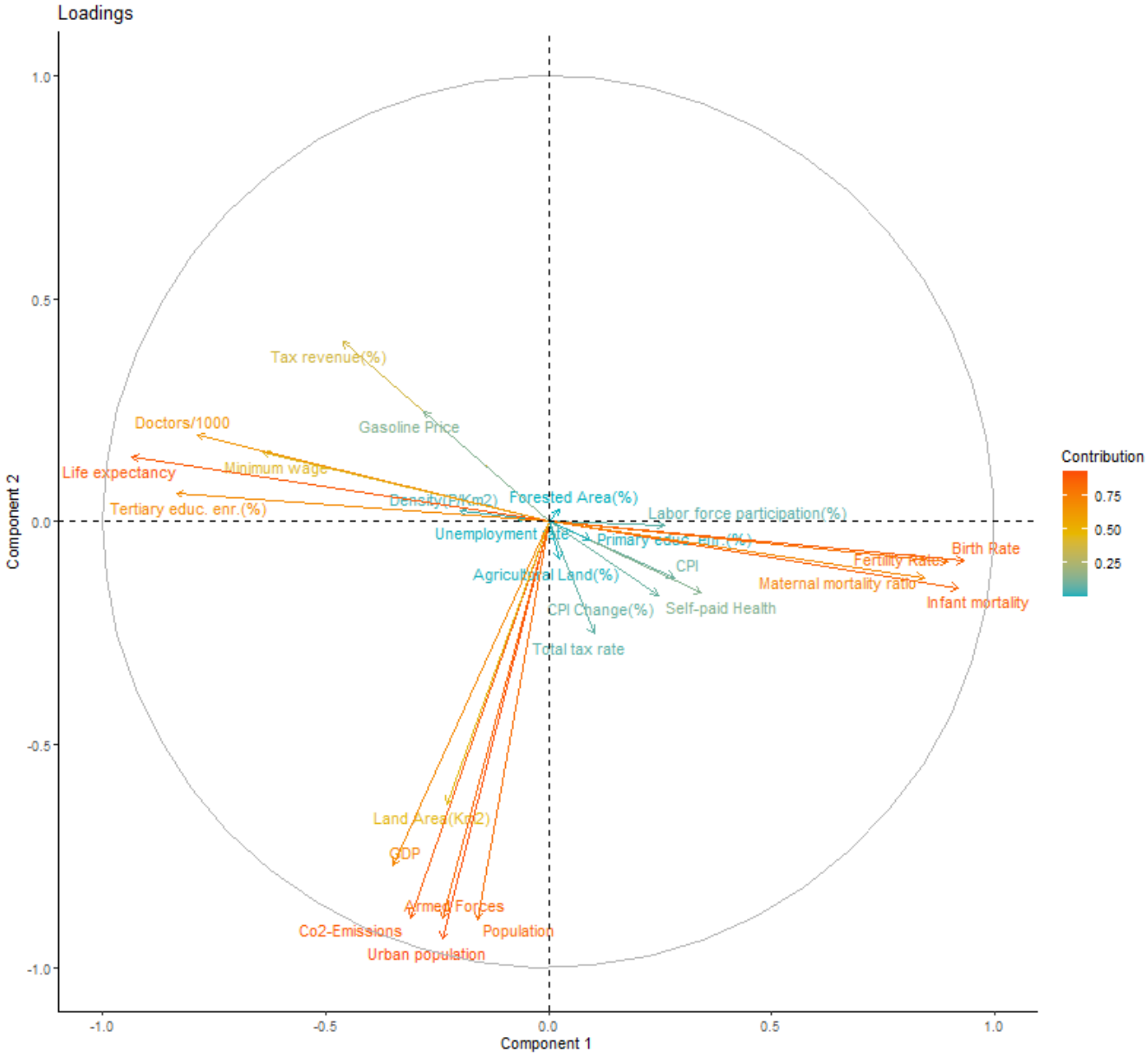


Figure 23: Biplot - loadings

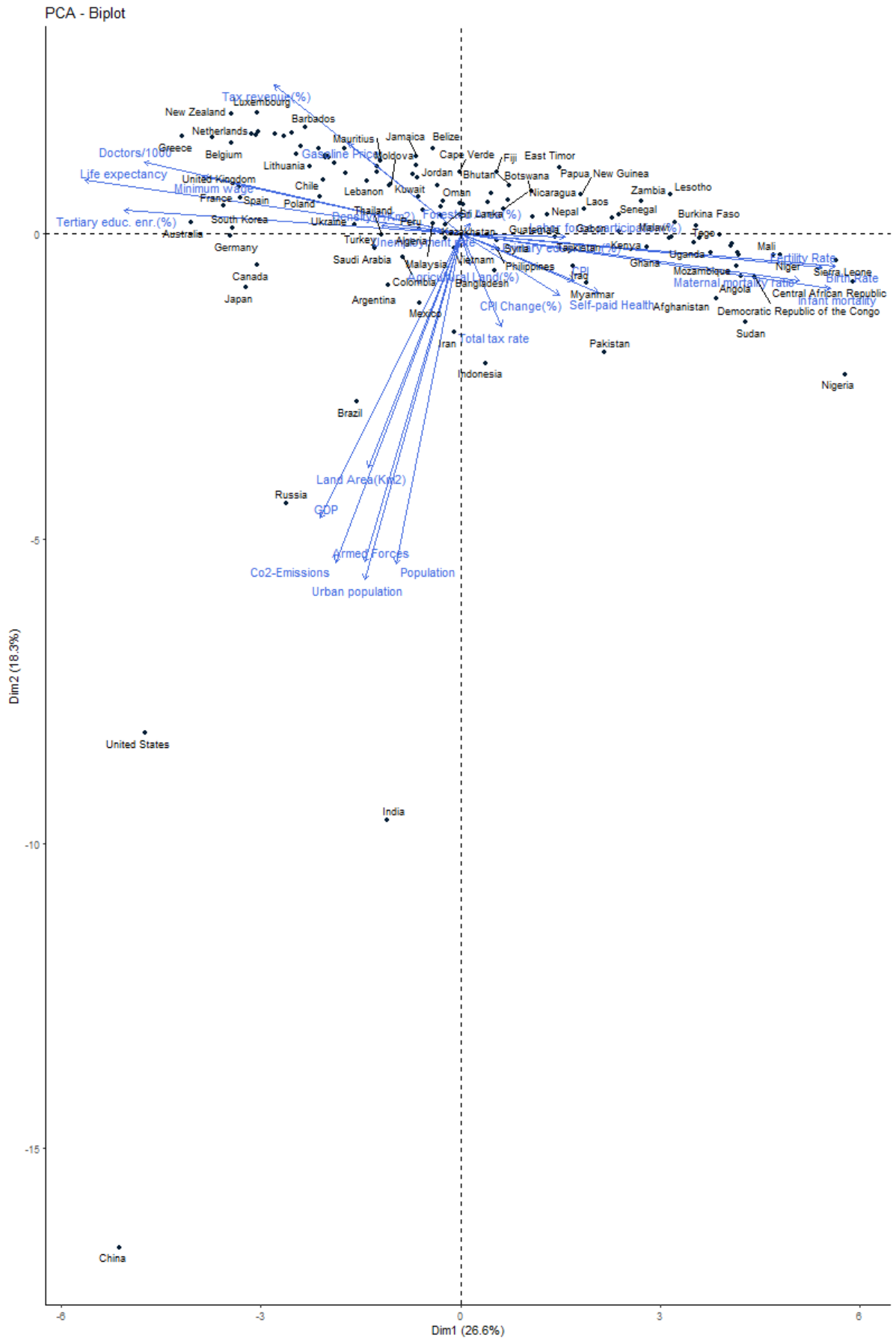


Figure 24: Biplot - loadings and scores

In the graph in figure 24 we then find the bi-plot with the corresponding scores in order to give an interpretation of how each country relates to each component. This type of plot helps visualize the structure of multivariate data, identifying clusters of countries and dominant contributing variables. Countries with high values along component 1 (e.g., Nigeria, Sudan, Niger, Central African Republic...) are likely to have high population growth, poor maternal/child health outcomes, and underdeveloped health systems while countries on the negative side (e.g., Germany, Australia, United Kingdom, France, South Korea...) show better public health and education systems. Countries like China, India, United States, Russia, and Brazil fall on the lower end of component 2 indicating large, populous, economically powerful nations with a significant carbon footprint. On the opposite side countries like Luxembourg, Netherlands, New Zealand appear higher up, associated with more fiscal policy-driven indicators and possibly environmental regulation.

3.3 Performing PCA with three components

Let's now perform a more elaborate PCA considering 3 components instead of two. In this case the total explained variance rises to 53.68% as can be seen from figure 21. In figure 25 we also find the relationships between each variable in the dataset and the three components taken into consideration with the respective communality.

	Component 1	Component 2	Component 3	Communality
Density(P/Km2)	-0.197	0.023	0.069	0.044
Urban population	-0.239	-0.936	-0.060	0.937
Population	-0.160	-0.894	-0.024	0.825
Armed Forces	-0.238	-0.889	0.025	0.848
GDP	-0.349	-0.770	-0.142	0.735
Minimum wage	-0.640	0.155	-0.151	0.456
Labor force participation(%)	0.259	-0.007	-0.649	0.488
Unemployment rate	-0.072	0.009	0.507	0.262
Tax revenue(%)	-0.462	0.403	-0.017	0.376
Total tax rate	0.102	-0.251	0.171	0.103
CPI	0.282	-0.126	0.532	0.378
CPI Change(%)	0.245	-0.165	0.550	0.390
Doctors/1000	-0.787	0.194	0.157	0.682
Self-paid Health	0.343	-0.160	0.441	0.338
Birth Rate	0.931	-0.088	-0.046	0.877
Fertility Rate	0.896	-0.091	-0.060	0.815
Life expectancy	-0.934	0.145	0.026	0.894
Infant mortality	0.919	-0.148	-0.054	0.869
Maternal mortality ratio	0.842	-0.125	-0.074	0.730
Co2-Emissions	-0.310	-0.889	-0.109	0.898
Land Area(Km2)	-0.230	-0.633	-0.068	0.458
Forested Area(%)	0.024	0.028	-0.639	0.410
Agricultural Land(%)	0.025	-0.085	0.384	0.155
Gasoline Price	-0.279	0.247	-0.184	0.173
Primary educ. enr.(%)	0.093	-0.040	-0.290	0.094
Tertiary educ. enr.(%)	-0.835	0.063	0.135	0.719

Figure 25: First three components and communality

Comparing these results with those obtained in figure 22 we can see that in some variables the communality, or how well the initial variables are explained by the components, increases. A significant example is the variable 'Labor force participation (%)', which goes from 0.067 to 0.488, therefore from 6.7% to 48.8%. Another variable involved in this increase is 'Forested Area (%)', which goes from 0.1% to 41%. These are the variables most influenced by the addition of this new component, which however explains only 8.8% of the total variance. Therefore, its addition

does not significantly change the analyses previously carried out with two components, but this has allowed us to understand how other variables that are not correlated with the first two components may be correlated with other components not included in the model. Obviously we could further investigate the PCA by adding a fourth component, but we would lose the graphic reproduction as it is not possible to have a graph with 4 dimensions.

We therefore proceed with the graphical analysis of the PCA composed of 3 components, which we find at the following link(3D graph loadings). From the graph, it is evident that the variables Forested Area and Labor Force Participation are negatively correlated with the third principal component, while Unemployment Rate, Consumer Price Index (CPI), CPI Change, Self-paid Health Expenditure, and Agricultural Land are positively correlated with it. This pattern suggests that the third component captures a contrast between factors associated with ecological and economic welfare, such as forest coverage and active labor market participation, and indicators of socio-economic pressure or territorial vulnerability, including high unemployment, inflation, healthcare costs borne by individuals, and extensive use of agricultural land. Therefore, this component could be interpreted as an index of social and environmental problems. Higher values of this component may reflect areas experiencing economic adversity and environmental stress, characterized by lower labor force engagement and reduced forested areas, while lower values might indicate regions with better ecological balance and stronger labor participation. In this sense, the third component provides insight into the degree of socio-economic and environmental vulnerability across territories.

At the following link(3D graph loadings+scores) instead we find the 3D graph which also contains the PCA scores, that is our countries and how these are distributed in relation to each component. Countries located on the positive side of the first component, such as Germany, France, Belgium, USA, Canada and the Netherlands, appear to be characterized by strong socioeconomic indicators, such as high GDP and robust education systems. These countries likely perform well across dimensions of infrastructure, governance, and health. On the negative side of Component 1, we find countries such as Mali, Niger, Cameroon, Angola and Sierra Leone, which are likely marked by limited access to basic services, lower economic output, and weaker institutional capacity. This suggests that Component 1 largely captures a development or modernization gradient. Along Component 2, countries like China, the United States, Russia, India and Brazil are located on the positive end. These are large, industrialized economies with substantial energy consumption and higher carbon emissions, suggesting that this component may be associated with industrial and environmental loadings. Component 3 introduces a further layer of differentiation. Countries such as China, Canada, United States and Japan are placed in the lower range of this component, perhaps due to the fact that they are countries with high levels of active labor force and at the same time have a high percentage of territory covered by forests. In fact, at the opposite end of this third component we find desert countries and those that suffer more from economic and social problems. Among these we find Syria, Iran, Argentina and Tunisia for example. Therefore, we recognize countries that have been living in economic crisis for years, such as Argentina, or that have a very low percentage of territory covered by forests, such as Iran and Syria.

3.4 Conclusion on Principal Component Analyses

The Principal Component Analysis has proven to be an effective tool for reducing the complexity of a dataset composed of 26 variables across 120 countries. By jointly applying the three main criteria: explained variance, eigenvalues greater than one, and the scree plot we were able to identify the most appropriate number of components, achieving a simplified yet meaningful representation of the data. The analysis with two principal components revealed a first dimension that captures the contrast between countries with high birth and mortality rates and those with stronger healthcare and education systems, effectively distinguishing between more and less developed nations. The second component added an economic and demographic dimension, highlighting differences in GDP, population size, military presence, and Co2 emissions. Expanding the analysis to three components allowed us to capture an additional layer of variability, particularly related to socio-economic and environmental stress factors. The third component differentiated countries with higher unemployment, inflation, and lower forest coverage from those with greater ecological stability and higher labor participation. Overall, PCA has enabled us to uncover global patterns and regional clusters, offering a multidimensional understanding of development, economic power, and environmental conditions. Compared to the earlier visualizations based on selected variables, PCA provides a more objective and structured way to understand the data. While direct visualization helped highlight some obvious patterns and groupings, PCA confirms those patterns and adds more depth by showing hidden relationships between variables. It also reveals new groupings that were not as clear before, making the overall analysis more complete and reliable. For example, many Sub-Saharan African countries cluster together in the upper region of the 3D space, revealing similar development challenges. In contrast, Scandinavian and Western European countries align in another distinct cluster, reflecting their emphasis on environmental regulation, fiscal responsibility, and welfare provision. Meanwhile, emerging powers such as Brazil, India, and Indonesia position themselves somewhere in between, reflecting both growing economic capability and persistent structural issues

4 Multidimensional Scaling

We will now conclude our analysis using nonlinear projection methods. Our goal is to analyze how these algorithms differ from the linear PCA method applied in the previous section, providing a more in-depth view of our dataset and its structure. The dataset used remains the same as in the PCA analysis, consisting of 120 observations, each representing a country characterized by 26 continuous variables.

The techniques we will employ in this section are Multidimensional Scaling (MDS), Locally Linear Embedding (LLE), and Isometric Feature Mapping (ISOMAP). Unlike PCA, which seeks to maximize variance and linearly transform the data to a lower-dimensional space, these nonlinear methods are designed to capture complex, nonlinear relationships in the data.

- Multidimensional Scaling (MDS): MDS aims to preserve pairwise distances between points in the dataset while projecting them into a lower-dimensional space. The algorithm iteratively adjusts the coordinates of the points to minimize the stress function, a measure of the disparity between original and projected distances. MDS is particularly useful for visualizing datasets with inherent distance-based structures but may struggle with highly nonlinear data structures compared to other techniques.

- Isometric Feature Mapping (ISOMAP): ISOMAP extends MDS by computing geodesic distances instead of Euclidean distances, allowing it to capture nonlinear manifold structures more effectively. By constructing a neighborhood graph and calculating shortest paths between points, ISOMAP preserves the global manifold structure, making it more suitable for datasets with complex curvature.

- Locally Linear Embedding (LLE): LLE attempts to preserve local neighborhoods by reconstructing each data point as a linear combination of its nearest neighbors. The algorithm then finds a low-dimensional embedding that maintains these local relationships. The sensitivity to the number of neighbors is significant in LLE, as setting too few neighbors may lead to fragmented clusters, while too many can cause over-smoothing of the manifold structure.

For LLE and ISOMAP, we will also investigate how the projections change when the number of neighbors is varied from 5 to 10 and finally to 15, as these algorithms are particularly sensitive to this parameter. This analysis will allow us to assess when each algorithm is more or less stable under different neighbor settings, providing insights into their robustness and suitability for specific data structures.

4.1 MDS Method

The plot in Figure 26 displays a two-dimensional projection of countries obtained using the Multi-Dimensional Scaling (MDS) method. The x-axis is labeled 'Component 1' and the y-axis is labeled 'Component 2' representing the two dimensions derived by MDS to best preserve the original pairwise distances between the countries in a lower-dimensional space. Each point on the plot represents a country, and its position is determined by its coordinates along these two MDS components. The goal of MDS is to arrange the countries in this 2D space such that the distances between any two countries in the plot are as close as possible to their corresponding distances in the original high-dimensional dataset. Therefore, countries that are plotted closer to each other are more similar based on the underlying characteristics or features used in the analysis, while countries that are further apart are less similar. The spread of the points across the two components indicates the degree of dissimilarity between the countries in the dataset.

As can be seen, the distribution of countries is very similar to that created by the PCA in figure 24. We can see countries that are further away from the others, including the United States, India, Russia and China. Sudan also appears to be far from the main cluster and is positioned on the opposite side of the graph compared to the richest countries mentioned above. In general, proceeding from left to right along the first component the wealth of the countries tends to increase, as it happened for the PCA graph. Countries positioned on the negative end of component 1 — such as Nigeria, Sudan, Niger, and the Central African Republic — tend to exhibit characteristics associated with high population growth, poor maternal and child health outcomes, and underdeveloped health systems. In contrast, countries on the positive side of component 1 — including Germany, Australia, the United Kingdom, France, and South Korea — are generally characterized by stronger public health systems and more advanced education infrastructures. Along component 2, countries like China, India, the United States, Russia, and Brazil cluster toward the lower end, representing large, populous, and economically influential nations with substantial carbon footprints. On the opposite end, nations such as Luxembourg, the Netherlands and Malta are positioned higher, potentially reflecting stronger emphasis on fiscal policy measures and environmental regulation.

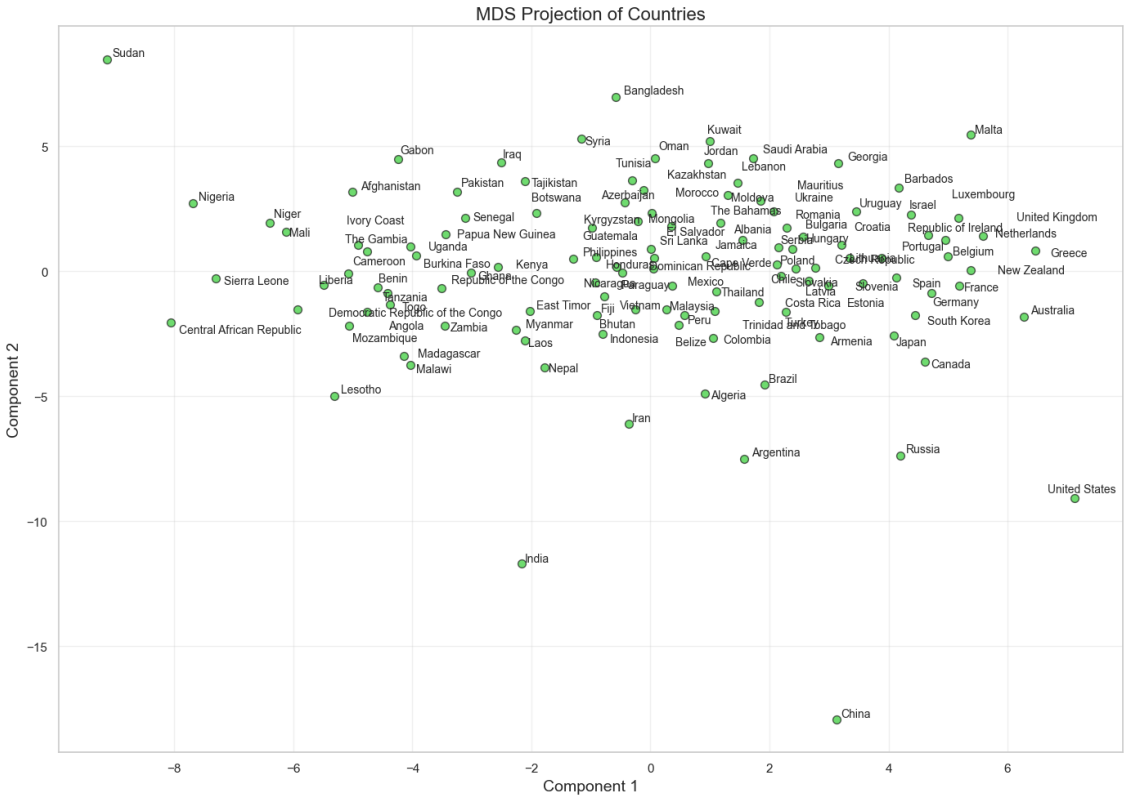


Figure 26: MDS Projection Method

4.2 ISOMAP Method

The sequence of ISOMAP projections, created using 5, 10, and 15 neighbors, demonstrates the sensitivity of manifold learning techniques to the definition of the local neighborhood. When ISOMAP operates with a small number of neighbors, such as 5 in Figure 27, it meticulously focuses on the immediate vicinity of each data point. This can result in a highly localized view of the data's structure forming distinct clusters of countries based on very specific similarities in the original high-dimensional space. For example, a clearer separation of the observations into two clusters can be noted at the value 5 of the first component compared to Figure 29 where 15 neighbors are considered.

As the number of neighbors is increased to 10 in the second image, ISOMAP begins to explore a broader context for each country. By considering a larger set of nearby points, the algorithm can start to bridge some of the gaps that might have existed with a smaller neighborhood size. This often leads to a more continuous and globally coherent embedding, where relationships between countries that are further apart in the local sense but connected through a series of intermediate neighbors become apparent. The overall structure of the data manifold starts to emerge more clearly. However, this expansion of the neighborhood also introduces a trade-off: the finer distinctions and local variations that were prominent with 5 neighbors might become somewhat smoothed out as the algorithm prioritizes capturing the larger-scale connectivity. In fact, it can be noted that the distinction into two clusters is less evident and outlier countries such as China, Sudan, Iran and Syria have come significantly closer to the countries that form the main cluster.

Finally, When ISOMAP employs 15 neighbors in Figure 29, it adopts a broader perspective, incorporating more distant points into each neighborhood. This can help reveal overarching patterns and relationships but also increases the risk of short-circuiting — a phenomenon where distant points are incorrectly brought closer together through expanded connections, distorting the true manifold structure. This effect is evident in the current visualization, where countries appear more closely packed, forming a more cohesive and dense cluster. Additionally, previously more isolated countries such as China, Russia, India, and the United States now appear noticeably closer to the main cluster, with the distance between these points and the central group significantly reduced compared to their positions in Figure 27. This convergence highlights how expanding the neighborhood size can smooth out the manifold, potentially obscuring distinct groupings while emphasizing more global connections.

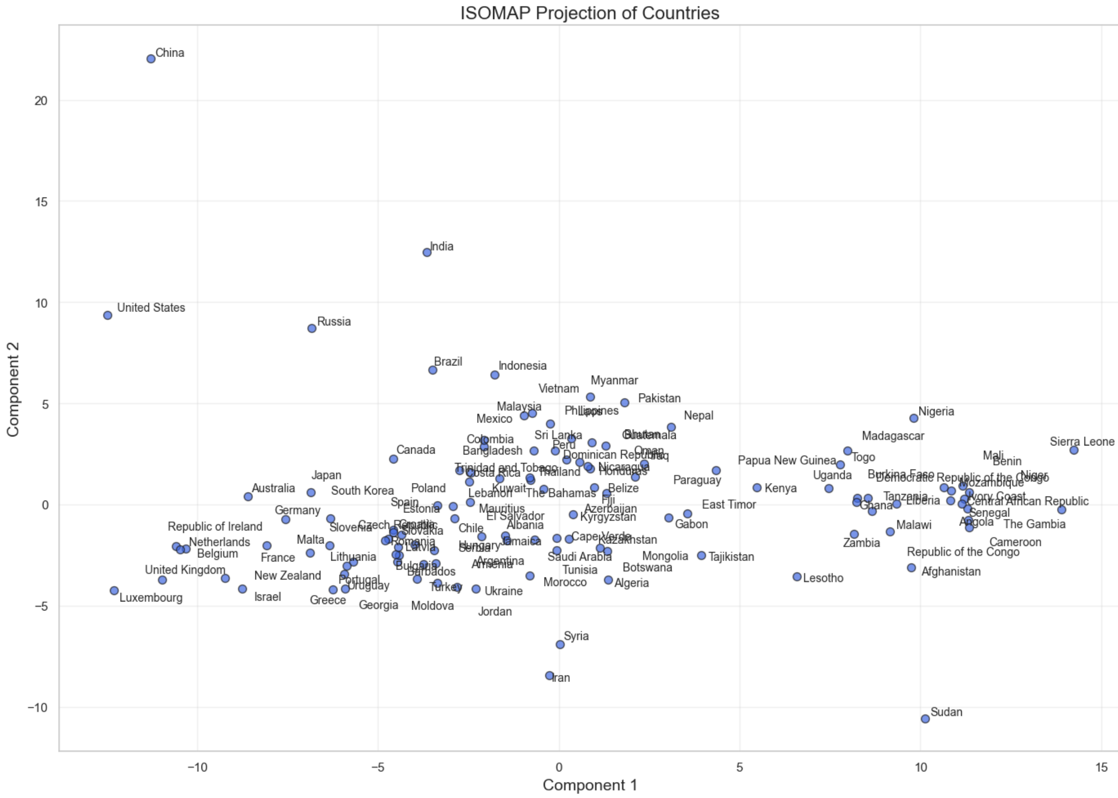


Figure 27: ISOMAP Projection Method with 5 Clusters

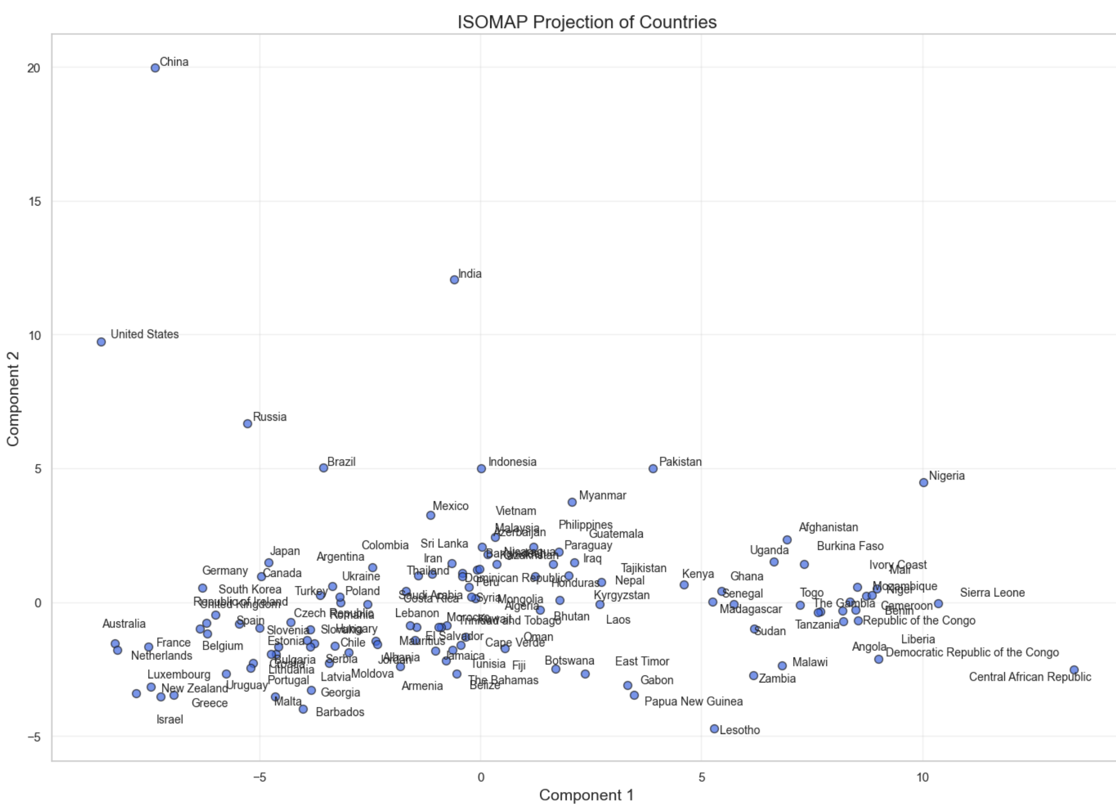


Figure 28: ISOMAP Projection Method with 10 Clusters

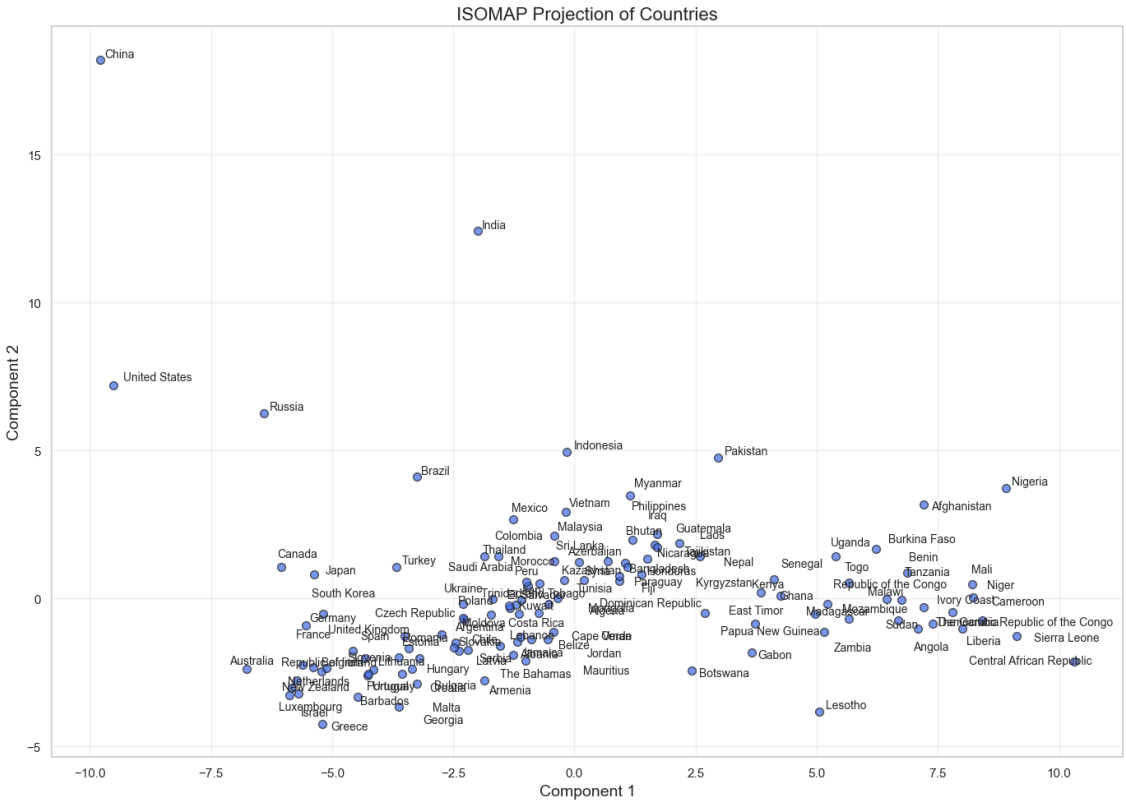


Figure 29: ISOMAP Projection Method with 15 Clusters

4.3 LLE Method

The Local Linear Embedding (LLE) projections, visualized with 5, 10, and 15 neighbors, illustrate a progressive transformation in the data's organization within the reduced two-dimensional space, emphasizing how neighborhood size shapes both local and global relationships. At the 5-neighbor setting, the plot reveals a somewhat fragmented structure, with clusters appearing tightly packed and certain groups of countries isolated. The overall shape of the data distribution is less defined, with more sharp transitions between clusters. This configuration highlights very local similarities but failing to resolve the broader, underlying manifold structure. Countries are grouped based on very specific, short-range relationships, resulting in a less cohesive global representation.

As the neighborhood size increases to 10, a noticeable smoothing effect emerges, creating greater connectivity among the data points. The clusters become less isolated, and more circular patterns start to form, indicating that LLE is beginning to capture slightly longer-range relationships. The overall shape of the projection becomes more organized, hinting at a more coherent representation of the data's intrinsic geometry. Transitions between clusters become more gradual, as the algorithm now considers a broader context for each country, producing a more interconnected layout that better reflects the data's manifold structure.

With 15 neighbors, the LLE projection adopts the most global perspective. The clusters merge more logical, and the overall shape of the embedding is more clearly defined, suggesting that the algorithm is now capturing even longer-range relationships and approximating the manifold structure with greater fidelity. However, this increased connectivity can come at a cost. While the global structure is better represented, some of the finer local distinctions observed with fewer neighbors may be lost. The algorithm's emphasis on broader connections can lead to over-smoothing, potentially bringing together countries that are only distantly related in the original high-dimensional space. We can see that using the LLE method the difference with the graph in Figure 30 is very marked. In the case in which 15 neighbors are considered, a graph very similar to the one seen previously for the MDS and PCA is obtained and we find the same arrangement of the countries, where China, the United States, Russia and India are outliers, distant from the central agglomeration of countries. Thus, the transition from 5 to 15 neighbors in LLE demonstrates a clear shift from a highly localized, fragmented representation to a more globally organized and continuous one. A smaller number of neighbors emphasizes local variations, whereas a larger number captures more extensive relationships, highlighting the underlying manifold structure.

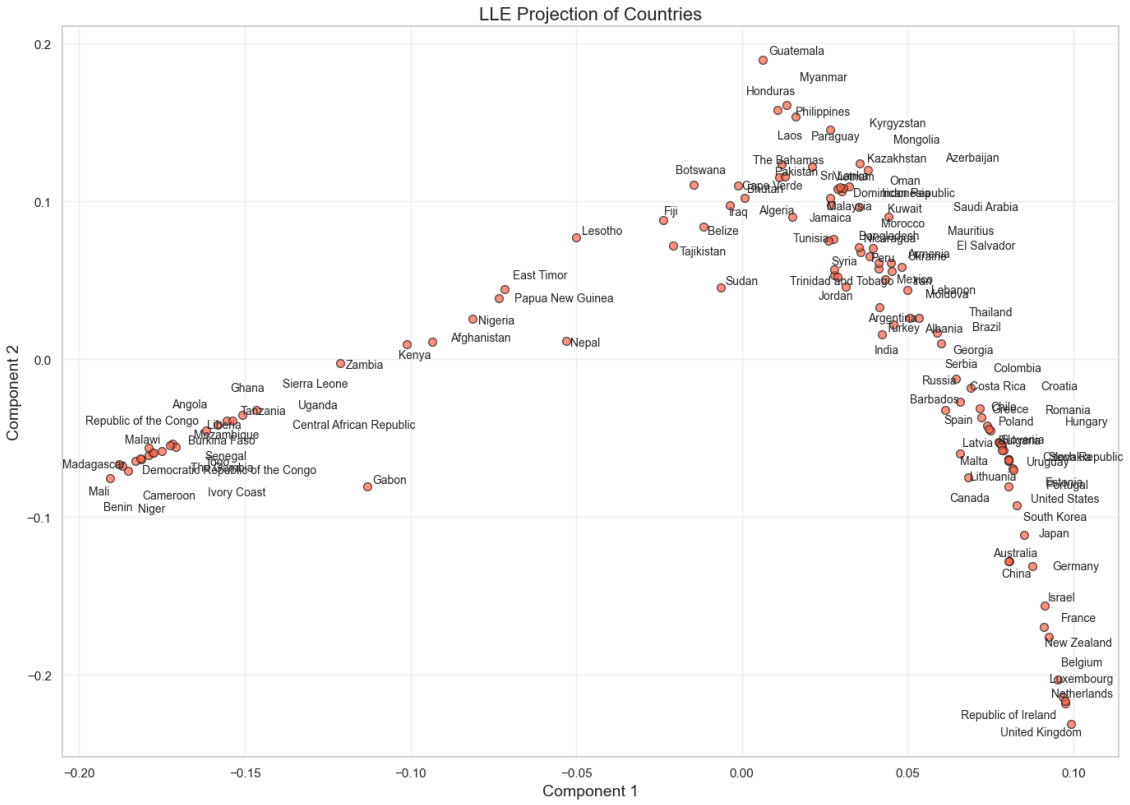


Figure 30: LLE Projection Method with 5 Clusters

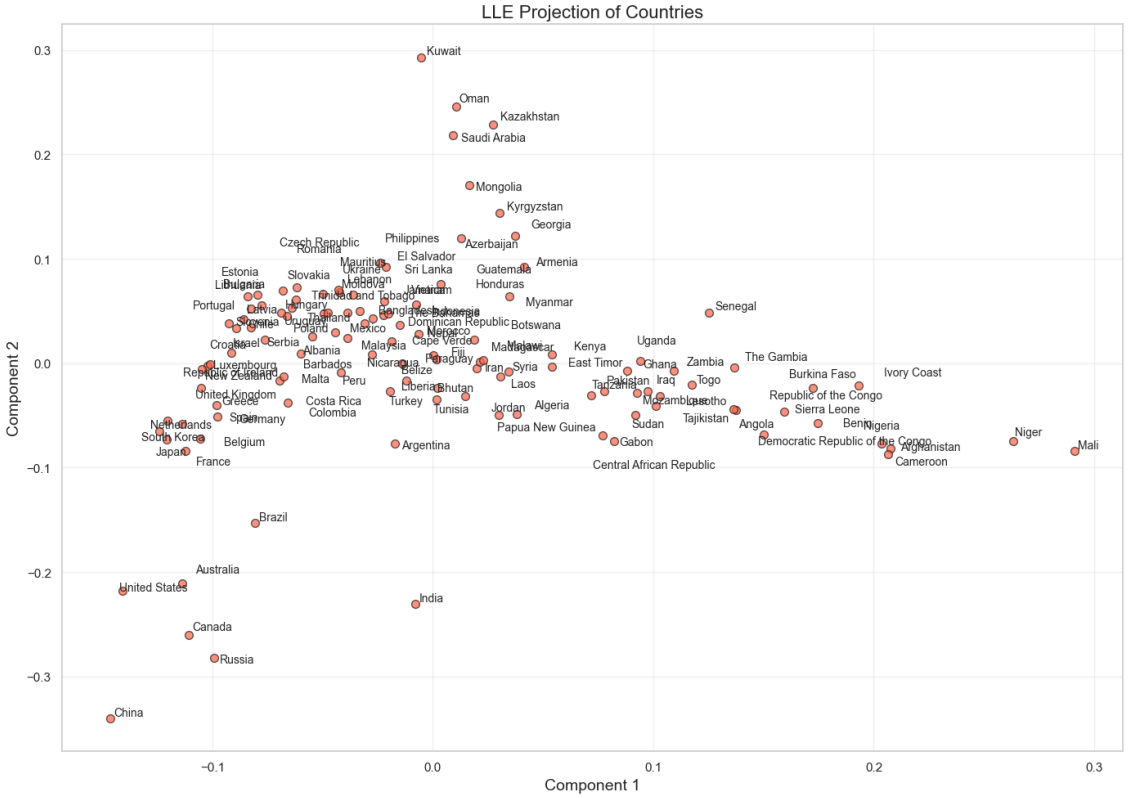


Figure 31: LLE Projection Method with 10 Clusters

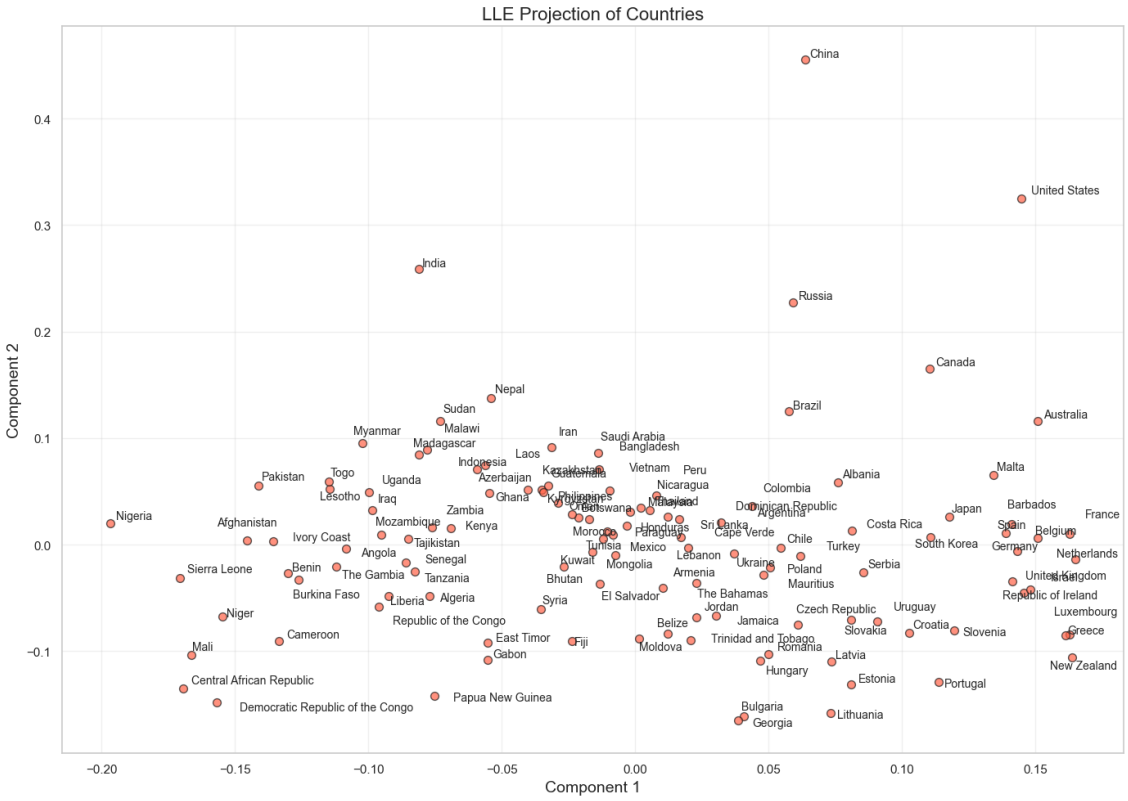


Figure 32: LLE Projection Method with 15 Clusters

5 General Conclusion

Applying both linear and nonlinear dimensionality reduction techniques gave us a clearer picture of the dataset of 120 countries and 26 variables. PCA did a solid job of simplifying the data by focusing on the main patterns related to economic development, health, and environmental factors. It helped us spot clear groupings, like Sub-Saharan African nations and Western European countries, based on things like economic power, population size, and carbon footprint. Interestingly, the two-component PCA plots ended up looking very similar to the projections from the nonlinear methods, showing that even a straightforward linear approach can capture the main structure of the data pretty well.

The nonlinear methods — MDS, ISOMAP, and LLE — took things a step further by uncovering more complex, nonlinear relationships. MDS stuck to preserving the pairwise distances but struggled a bit with more complicated structures, resulting in a more localized view. ISOMAP, on the other hand, did a better job of capturing the overall shape of the data by considering geodesic distances. However, as we increased the neighborhood size, it started to pull things closer together, sometimes distorting the actual relationships. LLE was all about local connections, and we saw how adjusting the neighborhood size could either keep the clusters tight and fragmented or smooth them out to form more connected patterns.

Despite the differences in approach, the projections from the nonlinear methods still looked similar to the PCA plots, especially when we used bigger neighborhood sizes in ISOMAP and LLE. This suggests that the main patterns PCA picked up are pretty solid and that even though non-linear methods can offer extra detail and context, the core structure of the data set is not all that different. In general, combining these techniques gave us a well-rounded view of the data, balancing the simplicity of PCA with the more intricate insights of non-linear methods.