Multidimensional Data Visualization - first exercise

Antonio De Patto

February 26, 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 Introduction

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.

2 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.

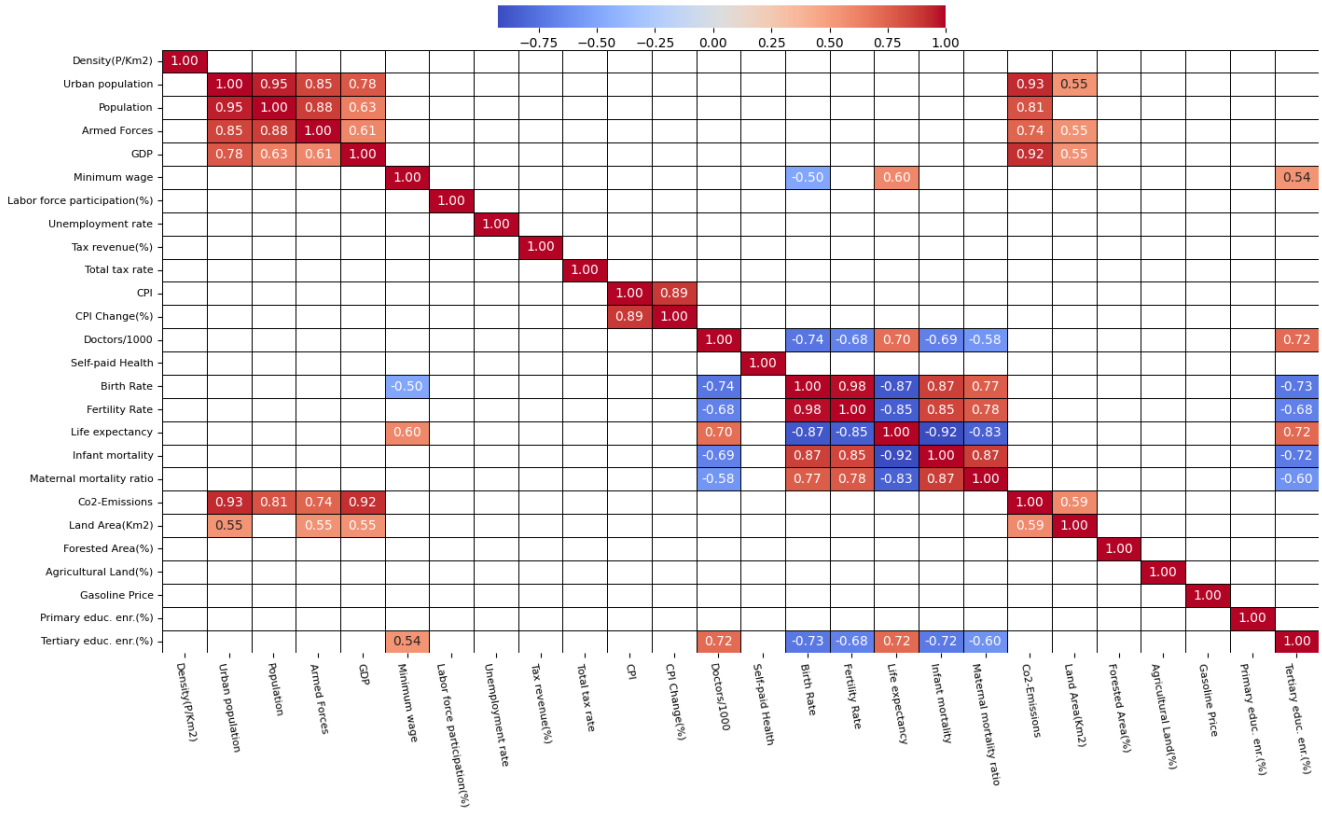

Figure 7: Correlation matrix

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.

2.1 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

2.2 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.

2.3 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.

2.4 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.

2.5 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.

3 Conclusion

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

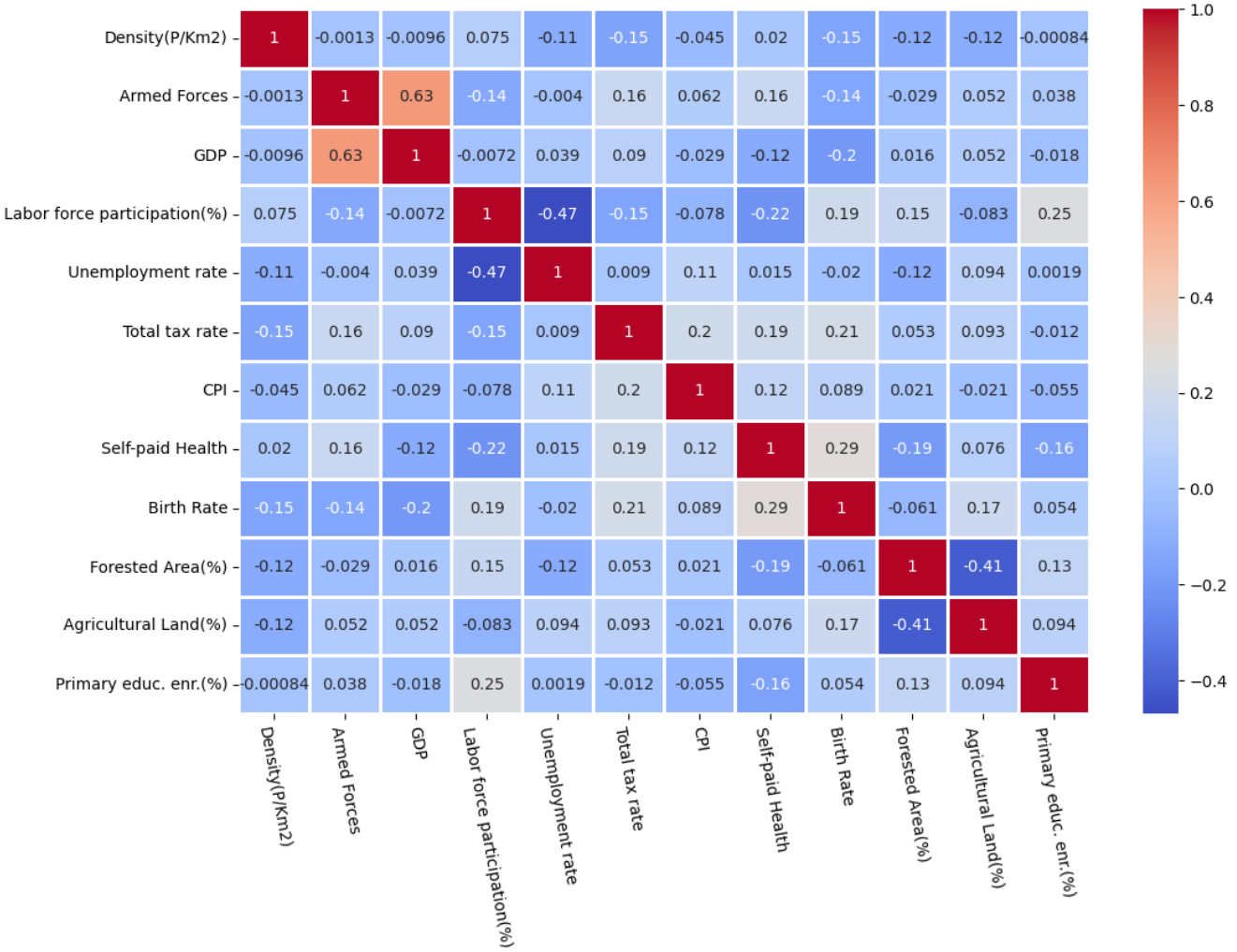

Figure 9: Correlation matrix of the final data set

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