Sustainable Development and Social WellBeing

A Multivariate Analysis of Economic, Environment and Social domains

Table of Contents

Introduction

The Multivariate Analysis of Economic, Social, and Environmental domains offer a comprehensive lens through which we examine the intricate interplay between economic factors, social trends, and environmental dynamics. This analytical approach enables a simultaneous exploration of multiple variables, uncovering relationships and patterns that might be overlooked in traditional single-variable analyses. Through an examination of the intricate interconnections among these three crucial domains, our aim is to acquire valuable insights into the comprehensive framework of societal development and sustainability for countries represented. Utilizing multivariate analysis approach, this report proves instrumental in understanding the multifaceted challenges and opportunities that arise at the intersection of economics, social, and the environmental domains , thereby contributing to a more informed and integrated approach to policy and decision-making. This report presents the results of a multivariate analysis of economic, environmental, and social data from 177 countries. The goal of the analysis was to provide insights into the relationships between various indicators of national development and social well-being and below are the various techniques that are used for the analysis:

1. Data Cleaning Techniques - This section describes the steps taken to preprocess the raw data, including handling missing values, converting non-numeric variables, checking for outliers, scaling variables. These steps ensured the data was cleaned and prepared for analysis.

2. Data Visualizations - PCA results and relationships between variables were visualized using plots like biplots and 3D scatterplots to aid interpretation. -Hemanth Kumar Ramanadham

3. Dimension Reduction Techniques - PCA and MDS were used to reduce the number of dimensions for analyzing patterns in the data. Results from PCA biplots ,scree plots, PC maps, MDS variable plots were discussed. -Mabel, Ogonna

4. Cluster Analysis - Techniques like hierarchical clustering, K-Means clustering, and model-based clustering were applied to group countries into meaningful clusters based on similarities.

- Yiming, Sun

5. Confirmatory Factor Analysis - Based on exploratory factor analysis, a conceptual model relating latent variables to observed variables was developed and tested using CFA. Model fit was evaluated. Combine all written reports and present them in a single report following all the report writing guidelines. -Leela Prasanna Akkala

1.Data Set

The primary dataset considered for our Multivariate analysis is Country Statistics - UN Data [1] .All the datasets that are available might not be suitable for multivariate analysis. When considering the multivariate data, the correlation between the variables is necessary, so the dataset is finalized considering the correlation values between the variables. To make the data ready for the analysis , the dataset was re-prepared by taking variables from different websites.

The finalized dataset contains a total of 20 variables and total observations of 177 Countries that are gathered from different source[2]. The table below shows the sample dataset with all the variables.

Country	Population	LE_Male	LE_Female	UnEmpRate	SexRatio	IMR	LitRate	CDI	GDP	PerCap	InfRate	Exports	Imports	Ecn_Agri	Ecn_Ind	Ecn_Serv	NDRI	NDRI_Cat	CO2	AQI	AQI_Cat	WQI	WQI_Cat
United States	338,289,857	76.1	81.1	3.6	96.9	5.4	99	85.69	25462700000000	65850	8.2	2.5 trillion	3.0 trillion	1.9	20.1	78	5.76	High	16.14	53	moderately polluted	86	Good
China	1,425,887,337	74.3	77.8	5.5	104.9	7.5	96.2	28.13	17963200000000	11400	2.5	2.6 trillion	2.1 trillion	9.2	41.1	49.7	7.7	Very High	7.38	62	unhealthy for sensitive groups	66	Poor
Japan	123,951,692	81	87.5	2.6	96.7	1.7	99	22.09	4231140000000	43100	3	683 billion	705 billion	1.2	26.4	72.4	5.73	High	9.49	37	good	85	Good
Germany	83,369,843	78.6	83.4	5.2	95.7	3.1	99	75.81	4072190000000	48000	8	1.5 trillion	1.2 trillion	0.6	30.5	68.9	3.91	Moderate	8.52	35	good	95	Excellent
India	1,417,173,173	69.7	71.8	7.6	107.9	27.3	77.7	80.98	3385090000000	2650	7.4	670 billion	680 billion	17	29.7	53.2	7.2	Very High	1.91	71	unhealthy	55	Very poor
United Kingdom	67,508,936	79.4	83.1	3.8	95.7	3.9	99	88.61	3070670000000	43300	10.1	452 billion	790 billion	0.7	19.4	79.9	4.21	Moderate	5.44	32	good	95	Excellent
France	64,626,628	79.5	85.6	7.4	95.3	3.1	99	86.35	2782910000000	42300	5.8	683 billion	780 billion	1.7	19.5	78.8	4.17	Moderate	4.83	36	good	93	Excellent
Russia	144,713,314	68	77.9	4	85.2	7.3	99.7	41.23	2240420000000	11150	15.6	416 billion	288 billion	4.9	38.9	56.2	4.44	Moderate	12.01	56	moderately polluted	83	Good
Canada	38,454,327	80.3	84	5.2	98.9	4.3	99	85.49	2139840000000	46500	6.9	523 billion	576 billion	1.8	28.6	69.6	4.49	Moderate	15.53	25	good	94	Excellent
Italy	59,037,474	80.3	85.1	8	94.7	2.5	98.2	73.86	2010430000000	37100	8	564 billion	544 billion	2.2	23.5	74.2	4.31	Moderate	6.19	46	moderate	91	Excellent
Brazil	215,313,498	72.8	79.6	9.3	97.2	16	92.6	74.05	1920100000000	13300	12	277 billion	256 billion	5.2	22.7	72	4.31	Moderate	2.41	52	moderate	74	Moderate

Figure 1.1 : Sample Dataset with the finalized variables.

Social	Population, LE_Male, LE_Female, UnEmpRate, SexRatio, IMR, LitRate, CDI.
Economic	GDP, PerCap, InfRate, Exports, Imports, Ecn_Agri, Ecn_Ind, Ecn_Serv
Environment	NDRI, NDRI_Cat, CO2, AQI, AQI_Cat, WQI, WQI_Cat.

Table 1.2 : Table to summarize the variables into different domains.

From the sample dataset it can be inferred that this analysis will be based on Economic, Environmental, and social aspects of various countries.

2. Multivariate Analysis

Multivariate Analysis refers to the set of techniques that are used to analyze multivariate data to understand their relationships and patterns. The main goal is to explore the data and understand it.

There are several techniques that are available for multivariate Analysis. We will discuss the techniques that are used in our project.

3. Data Cleaning Techniques

Data Cleaning and Preprocessing is the crucial step in the implementation of multivariate analysis on a given dataset. The process involves identifying any missing values and handling them with appropriate techniques like the Value Insertion methods, pairwise or listwise deletion techniques, and maximum likelihood estimations. Data cleaning also involves identification of the outliers and removing any data inconsistencies. It is important to make sure that all the variables are in the same direction and appropriate techniques like scaling, dividing by range should be implemented. All these measures ensure that data is thoroughly cleaned and is available for analysis.

In our project, there are several steps that are involved as part of data cleaning that are described below:

3.1. Adding of new Columns to the dataset:

The first step involves addition of a new column or the variable to the dataset continents. As we were dealing with the countries data we grouped the data in continents , from this we wanted to handle the missing values more effectively. The main reason for adding the continent column was to impute missing values at the country level with its continent estimates.

The new column was added by using a function to group all countries by their continent. However, during testing we noticed some issues:

- The function was unable to identify the continent for Micronesia. This would result in missing values being inserted in the continent column for this country.
- All North and South and Central American countries were grouped as "Americas". For our analysis, it was necessary to separate these regions into distinct continents/subcontinents.

Therefore, we manually hardcoded these cases to improve the accuracy and granularity of the continent classification. Specifically, we hard coded Micronesia to the correct continent of "Oceania". We also separated North America, South America, and Central America into individual categories rather than lumping them together as "America."

3.2.Handling Missing Values :

There are some missing values in the dataset, and we used the Median imputation method for handling the missing values. Median values are commonly used for imputation as they are not influenced by outliers and represent the central tendency of the continent. Therefore, we chose to impute any missing country values with the median of the respective continent.

3.3. Numeric Conversion:

The dataset contains two variables namely Exports and Imports which contain the alpha numeric data, so aiming to convert this data to numerical data , we implemented a custom function to convert export and import values denoted in trillion, billion, and million to numeric format. This ensures consistency and accuracy in the representation of these values for further analysis. The Export and Import values were provided with label suffixes like "trillion", "billion" and "million" to indicate the magnitude.

3.4. Direction changing:

Another important step in data cleaning is making sure that all the variables are in the same scale. As we have few variables that have different directions, we employed the dividing by range technique to change the direction of certain variables. This reversal was applied to enhance the interpretability of the data, particularly for indicators where lower values traditionally represent more favorable conditions.

3.5. Detecting Outliers:

Detecting Outliers is also crucial step but While outliers were present in the data for countries like United States , China, India which are with high population and some countries with lower population that we have deliberately opted not to remove. Removing outliers could discard important information and introduce bias against large economies. It is preferable to retain the entire range of values for analysis rather than filtering out potential outliers.

3.6. Scaling and Preparing the final data.

While reading the data, Country variable to set to the row names to enhances the clarity and accessibility of the dataset, aligning row identifiers with the corresponding countries. All the numeric variables are considered, and the categorical variables are neglected purposely. The data is scaled and correlation between the variables is identified to ensure an appropriate dataset before the analysis.

	Population	LE_Male	LE_Female	UnEmpRate	SexRatio	IMR	LitRate	CDI	GDP	PerCap	InfRate	Exports	Imports
United States	1.89	0.70	0.74	0.58	-0.0068	0.77	0.72	0.96	10.4	2.35	0.18	7.4	9.0
China	8.86	0.46	0.32	0.29	0.4326	0.66	0.56	-3.93	7.2	-0.24	0.38	7.7	6.2
Japan	0.51	1.35	1.55	0.73	-0.0178	0.96	0.72	-4.44	1.5	1.27	0.36	1.8	1.9
Germany	0.25	1.03	1.03	0.34	-0.0727	0.89	0.72	0.12	1.5	1.50	0.19	4.3	3.4
India	8.80	-0.15	-0.44	-0.02	0.5973	-0.34	-0.51	0.56	1.2	-0.65	0.21	1.7	1.8
United Kingdom	0.15	1.14	0.99	0.55	-0.0727	0.85	0.72	1.20	1.0	1.28	0.11	1.1	2.1

	Ecn_Agri	Ecn_Ind	Ecn_Serv	NDRI	CO2	AQI	WQI
United States	-0.83	-0.64	1.25	0.33	-2.30	-0.43	0.63
China	-0.23	1.07	-0.72	1.56	-0.58	-0.86	-0.80
Japan	-0.88	-0.13	0.86	0.32	-0.99	0.35	0.56
Germany	-0.93	0.21	0.62	-0.84	-0.80	0.44	1.27
India	0.40	0.14	-0.47	1.25	0.49	-1.30	-1.58
United Kingdom	-0.92	-0.70	1.38	-0.65	-0.20	0.59	1.27

Figure 3.1 : Top 6 observations of the scaled dataset.

	Population	LE_Male	LE_Female	UnEmpRate	SexRatio	IMR	LitRate	CDI	GDP	PerCap	InfRate	Exports	Imports
Population	1.00	0.00	-0.02	0.02	0.06	0.00	-0.02	-0.25	0.58	-0.05	0.01	0.62	0.56
LE_Male	0.00	1.00	0.96	0.12	-0.04	0.87	0.72	-0.08	0.15	0.66	0.11	0.28	0.28
LE_Female	-0.02	0.96	1.00	0.07	-0.06	0.92	0.76	-0.11	0.15	0.63	0.09	0.27	0.27
UnEmpRate	0.02	0.12	0.07	1.00	0.11	0.03	-0.15	0.08	0.08	0.18	0.17	0.12	0.12
SexRatio	0.06	-0.04	-0.06	0.11	1.00	-0.09	-0.09	0.21	0.01	0.01	-0.01	0.00	0.01
IMR	0.00	0.87	0.92	0.03	-0.09	1.00	0.80	-0.13	0.15	0.52	0.07	0.24	0.24
LitRate	-0.02	0.72	0.76	-0.15	-0.09	0.80	1.00	-0.12	0.13	0.45	0.03	0.20	0.19
CDI	-0.25	-0.08	-0.11	0.08	0.21	-0.13	-0.12	1.00	-0.16	0.12	0.14	-0.19	-0.12
GDP	0.58	0.15	0.15	0.08	0.01	0.15	0.13	-0.16	1.00	0.22	0.04	0.91	0.94
PerCap	-0.05	0.66	0.63	0.18	0.01	0.52	0.45	0.12	0.22	1.00	0.13	0.34	0.34
InfRate	0.01	0.11	0.09	0.17	-0.01	0.07	0.03	0.14	0.04	0.13	1.00	0.06	0.06
Exports	0.62	0.28	0.27	0.12	0.00	0.24	0.20	-0.19	0.91	0.34	0.06	1.00	0.98
Imports	0.56	0.28	0.27	0.12	0.01	0.24	0.19	-0.12	0.94	0.34	0.06	0.98	1.00
Ecn_Agri	0.03	-0.64	-0.68	0.19	0.09	-0.68	-0.70	0.10	-0.13	-0.54	-0.13	-0.22	-0.22
Ecn_Ind	0.07	-0.06	-0.09	-0.09	-0.12	-0.07	0.06	-0.35	0.02	-0.03	0.03	0.05	0.01
Ecn_Serv	-0.08	0.60	0.66	-0.09	0.03	0.64	0.55	0.21	0.10	0.49	0.09	0.15	0.17
NDRI	0.22	-0.75	-0.76	0.08	0.11	-0.70	-0.69	0.13	0.05	-0.64	-0.15	-0.05	-0.04
CO2	-0.01	-0.53	-0.51	-0.11	0.10	-0.50	-0.46	0.22	-0.23	-0.56	-0.11	-0.30	-0.29
AQI	-0.23	0.51	0.55	0.02	-0.01	0.54	0.49	0.38	-0.06	0.46	0.12	0.00	0.00
WQI	-0.18	0.87	0.89	0.00	-0.10	0.84	0.78	-0.05	0.07	0.69	0.11	0.18	0.18

	Ecn_Agri	Ecn_Ind	Ecn_Serv	NDRI	CO2	AQI	WQI
Population	0.03	0.07	-0.08	0.22	-0.01	-0.23	-0.18
LE_Male	-0.64	-0.06	0.60	-0.75	-0.53	0.51	0.87
LE_Female	-0.68	-0.09	0.66	-0.76	-0.51	0.55	0.89
UnEmpRate	0.19	-0.09	-0.09	0.08	-0.11	0.02	0.00
SexRatio	0.09	-0.12	0.03	0.11	0.10	-0.01	-0.10
IMR	-0.68	-0.07	0.64	-0.70	-0.50	0.54	0.84
LitRate	-0.70	0.06	0.55	-0.69	-0.46	0.49	0.78
CDI	0.10	-0.35	0.21	0.13	0.22	0.38	-0.05
GDP	-0.13	0.02	0.10	0.05	-0.23	-0.06	0.07
PerCap	-0.54	-0.03	0.49	-0.64	-0.56	0.46	0.69
InfRate	-0.13	0.03	0.09	-0.15	-0.11	0.12	0.11
Exports	-0.22	0.05	0.15	-0.05	-0.30	0.00	0.18
Imports	-0.22	0.01	0.17	-0.04	-0.29	0.00	0.18
Ecn_Agri	1.00	-0.32	-0.59	0.71	0.52	-0.43	-0.75
Ecn_Ind	-0.32	1.00	-0.58	-0.16	-0.18	-0.30	-0.01
Ecn_Serv	-0.59	-0.58	1.00	-0.48	-0.29	0.62	0.65
NDRI	0.71	-0.16	-0.48	1.00	0.57	-0.46	-0.87
CO2	0.52	-0.18	-0.29	0.57	1.00	-0.11	-0.59
AQI	-0.43	-0.30	0.62	-0.46	-0.11	1.00	0.63
WQI	-0.75	-0.01	0.65	-0.87	-0.59	0.63	1.00

Figure 3.2. The correlation matrix of all the variables of the dataset

4. Data Visualizations

Data visualization transforms complex information into easily understandable visual representations, facilitating insights and informed decision-making. Data visualization is crucial because it translates large and complex datasets into easily understandable visual representations, enabling quicker insights, pattern recognition, and effective communication of information to support informed decision-making in various fields such as business, science, and public policy.

We wanted to group other all the visualizations together. So Mahalanobis distances plot which is used for outlier detection is also included in here.

4.1. Mahalanobis distances

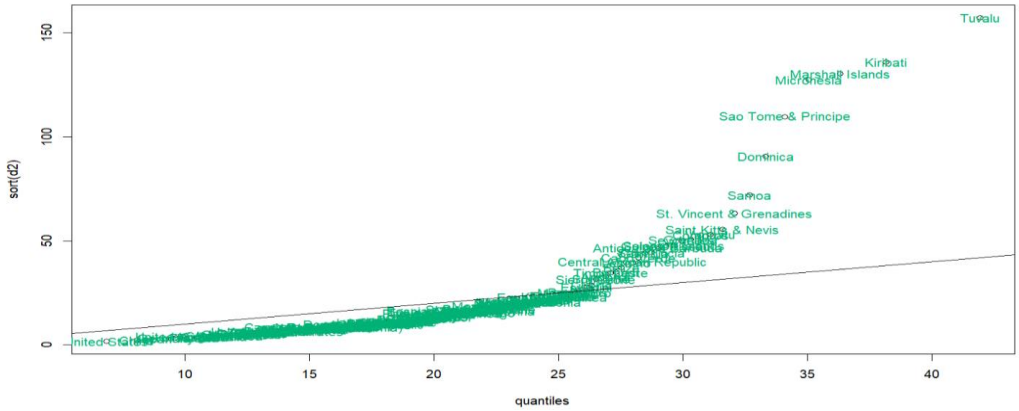

Figure 4.1 : Mahalanobis distance – outlier detection

4.2. Box Plot – Univariate

Boxplot can also be another outlier detection technique that can be applied for a univariate variable. This is a sample boxplot and can be applied to every variable.

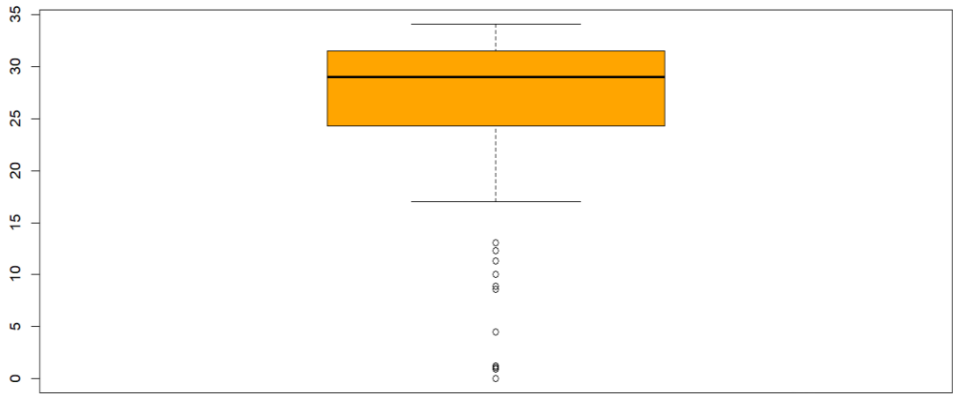

Figure 4.2 : Univariate Boxplot – UnEmpRate

4.3. Bi -Variate Box Plot

Bivariate Boxplot displays the distributional properties of the data and highlights the potential outliers. Like univariate boxplot this can be extended to all the pairs of variables to individually check the outliers. Below is a visual representation for WQI and AQI. From this we can infer that Kuwait, Bahama are the outliers.

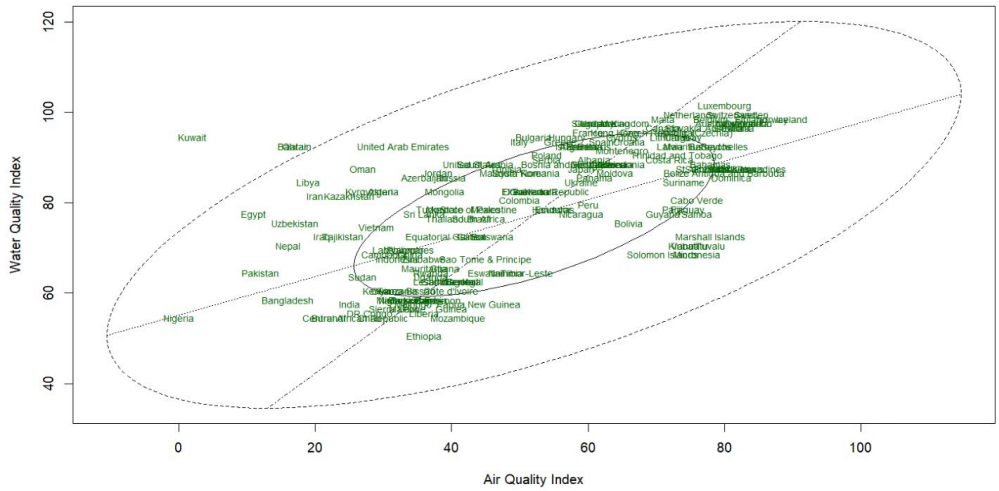

Figure 4.3 : Bivariate Boxplot – AQI vs WQI

4.4. Bubble Plot

Bubble plot can be used to include more than 2 variables in scatterplot and the third variable can be represented by radii of circles proportioned to values.

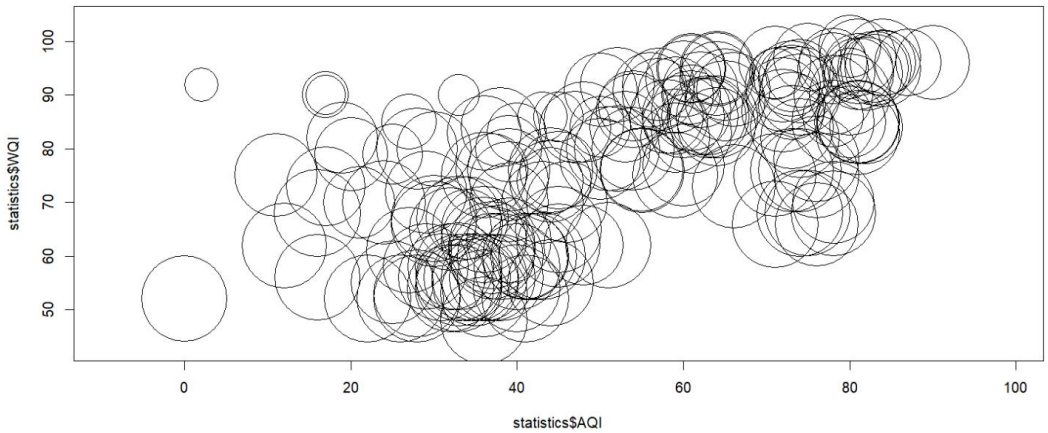

Figure 4.4 : Bubble plot for environmental data , AQI, WQI, CO2

4.5. Scatterplot Matrix

Scatterplots can be used to identify the patterns or the outliers and the correlation between the variables . Scatterplots can be applied on pair of variables or even the entire dataset variables in the form of scatterplot matrix. This is one of the most crucial visualization techniques to understand the data.

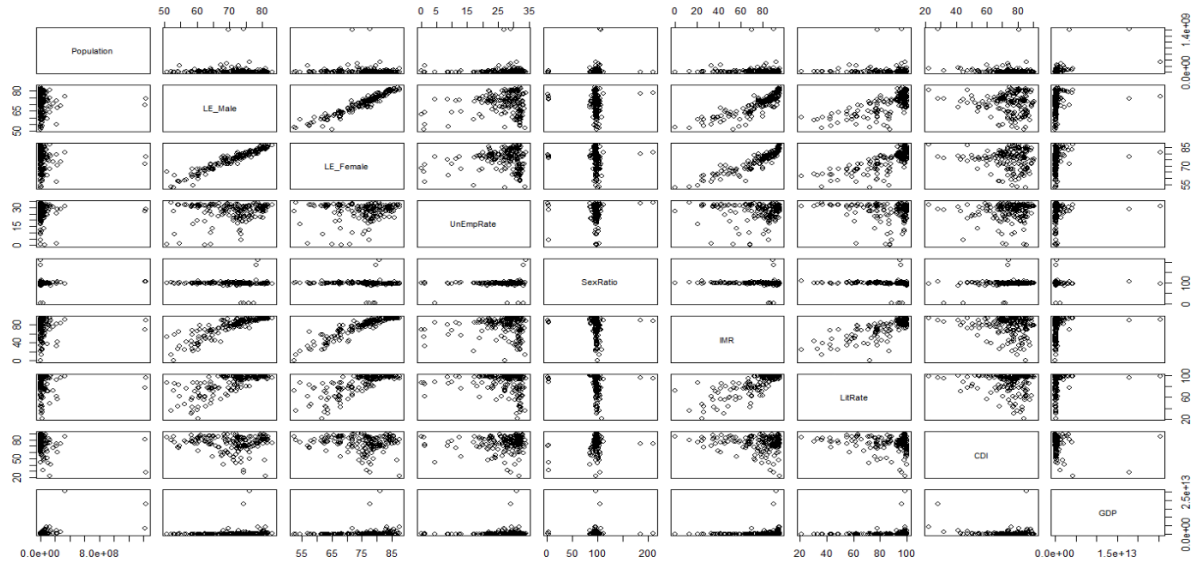

Figure 4.5. Scatterplot Matrix for the first 9 variables in the dataset

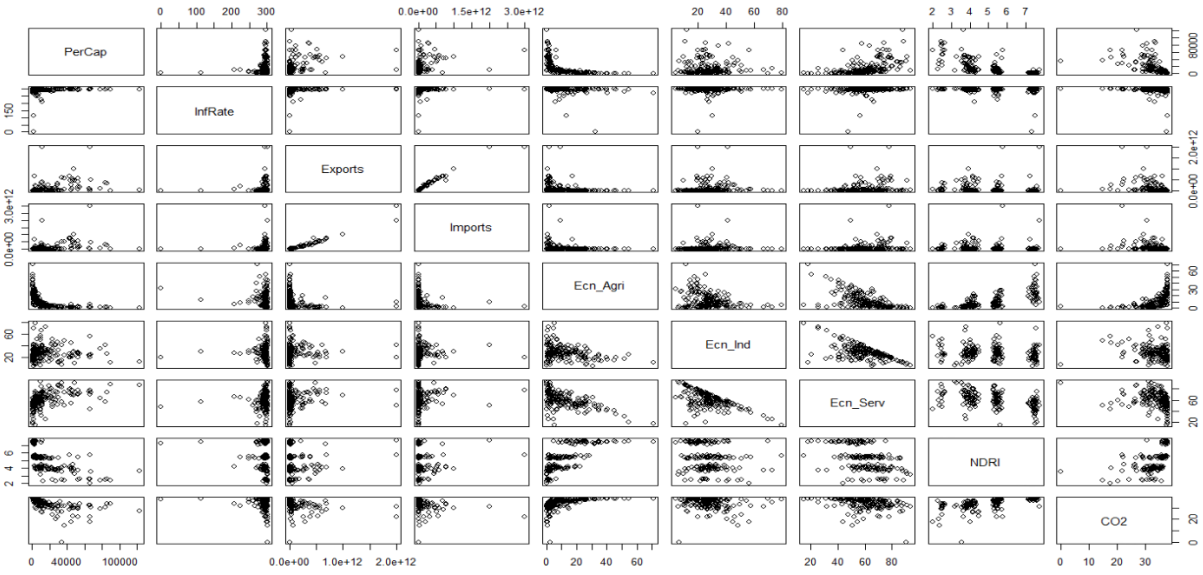

Figure 4.6. Scatterplot Matrix for the last 9 variables in the dataset

4.6. Scatterplot 3 D

A scatterplot can also be visualized in a 3-D format for a clearer visual representation. A sample 3-D scatterplot was plotted for Inflation rate , literacy rate and unemployment rate.

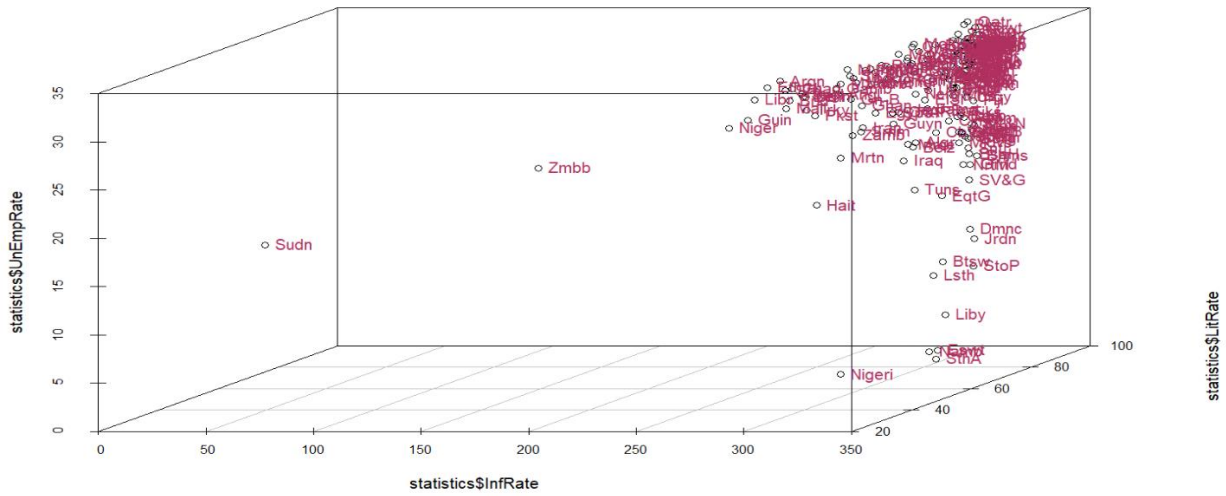

Figure 4.7. Scatterplot – 3D for UnEmpRate, InfRate, LitRate variables in the dataset.

4.7 GG Plot

GG plot can be used for various purposes, in our project we used gg plot to visually represent the population distribution of the countries which has a major impact on the socio-economic conditions.

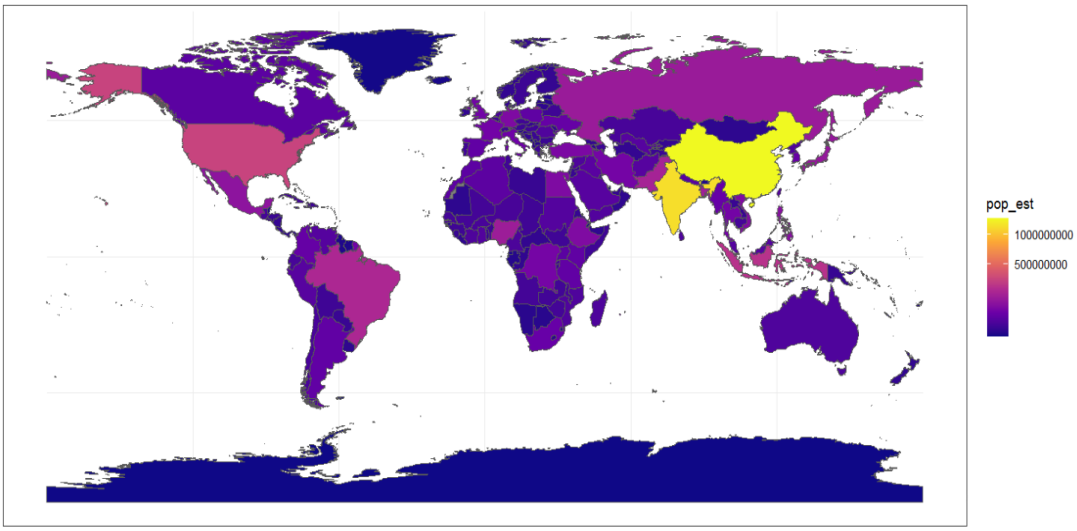

Figure 4.8: GGplot representing the population distribution on a map.

5. Dimension Reduction Techniques

Dimension reduction techniques involve reducing the number of variables so that at least 75 percent of the data is represented in the considered components. There are several dimension reduction techniques among them we have considered Principal Component analysis.

5.1. Principal Component Analysis:

PCA was used for dimensionality reduction and identifying patterns in multivariate data. We explored the importance of principal components and their loadings to gain insights into the underlying structure of the dataset.

5.2. Results of the Principal Component Analysis:

```
Importance of components:
                       Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9 Comp.10 Comp.11 Comp.12 Comp.13 Comp.14 Comp.15
Standard deviation      2.76   1.84   1.44  1.143   1.00  0.974   0.90  0.813  0.650   0.621   0.587   0.545   0.473  0.4196  0.3101
Proportion of Variance  0.38   0.17   0.10  0.066   0.05  0.048   0.04  0.033  0.021   0.019   0.017   0.015   0.011  0.0089  0.0048
Cumulative Proportion   0.38   0.55   0.66  0.722   0.77  0.820   0.86  0.894  0.915   0.934   0.952   0.967   0.978  0.9868  0.9916
                       Comp.16 Comp.17 Comp.18 Comp.19 Comp.20
Standard deviation      0.2731  0.2360  0.1558  0.1095 3.9e-03
Proportion of Variance  0.0038  0.0028  0.0012  0.0006 7.7e-07
Cumulative Proportion   0.9954  0.9982  0.9994  1.0000 1.0e+00

Loadings:
           Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9 Comp.10 Comp.11 Comp.12 Comp.13 Comp.14 Comp.15 Comp.16
Population         0.416          0.146                 0.359  0.114  0.765  0.125           0.152           0.148
LE_Male    -0.331                       -0.186         0.132                0.162  -0.265  -0.265  -0.212          -0.401   0.253
LE_Female  -0.339                       -0.174         0.158  0.104                -0.180  -0.226  -0.195  -0.134  -0.134
UnEmpRate                  0.190 -0.638 -0.463 -0.183         0.287         -0.248   0.304           0.161   0.189
SexRatio                   0.251 -0.127 -0.166  0.874  0.269 -0.124 -0.169
IMR        -0.325                       -0.153         0.213  0.152         -0.179  -0.155          -0.169  -0.235   0.735  -0.227
LitRate    -0.298         -0.108  0.180                 0.134  0.174         -0.348  -0.226   0.248   0.728           -0.167
CDI               -0.175  0.470 -0.127  0.356  0.150 -0.329  0.192  0.284 -0.283  -0.422  -0.250           0.167
GDP                0.484  0.119                       -0.118         -0.272                                  0.111   0.250   0.710
PerCap     -0.270                -0.255                -0.399 -0.139  0.218  0.510                   0.366  -0.425   0.175
InfRate                          -0.483  0.581 -0.209  0.547 -0.194 -0.131                                  -0.102
Exports    -0.133  0.481  0.103                                      -0.145                                          -0.165  -0.550
Imports    -0.132  0.474  0.135                       -0.147         -0.234                                                  -0.233
Ecn_Agri    0.291         0.186         -0.289 -0.187                -0.112  0.240  -0.529   0.318
Ecn_Ind                  -0.559 -0.232  0.242  0.271 -0.116  0.384                                  -0.110
Ecn_Serv   -0.259 -0.106  0.319  0.253                       -0.304         -0.149   0.402  -0.207
NDRI        0.300  0.135  0.190                                             -0.317                          -0.736  -0.252
CO2         0.231         0.196  0.246                 0.224  0.552 -0.227  0.427   0.145  -0.362   0.263
AQI        -0.220 -0.197  0.288         0.172                 0.406                 0.249   0.639  -0.302
WQI        -0.341 -0.102                                                                                  0.227  -0.223
           Comp.17 Comp.18 Comp.19 Comp.20
```

Figure 5.1: PCA Result

The PCA identified 5 principal components related to social, economic, and environmental factors that explained 77% of the total variance in the data.

The first principal component (Comp.1) explained the largest proportion of variance (38%) with positive loadings on IMR and Ecn_Agri. Negative loadings on Ecn_Serv, WQI, LE_Male and LE_Female suggest similarities for countries that perform poorly in these areas.

The combination of positive loadings on infant mortality, Natural Disaster Recovery Index, Air Quality Index, and negative loadings on life expectancy, literacy rate, per capita income, CO2

emissions, economic services, and water quality suggests that this component may represent a profile associated with challenges in healthcare, economic development, and environmental quality. This profile is more indicative of countries in Africa continent facing socio-economic and environmental challenges, high infant mortality rate and high reliance on Agriculture .

PC2 indicates a continent with a combination of large population, actively trade-focused, strong economic performance, better environmental quality, and lower cultural diversity compared to other regions. This profile is more indicative of developed and developing regions or continents that have experienced higher economic growth and stability such as North America, Europe and some Asian countries like China, Japan, Saudi Arabia, etc.

Countries in Oceania and Western Europe are best described by PC3 with the component capturing patterns around Cultural diversity and gender ratios, and an economy oriented more towards services rather than industry.

The profile of PC4 is indicative of regions with higher unemployment rates, higher inflation rates, and a higher contribution from economic services may best capture variations within some African and Caribbean countries.

Profiles with higher unemployment rates, lower sex ratio, higher cultural diversity, lower inflation rates, a lower contribution from the agricultural sector, and a higher contribution from the industrial sector suggests characteristics associated with a diverse and potentially industrialized region are loaded on PC5. Listed island nations and parts of Africa align as vulnerable to global trade shifts captured here.

5.3. PCA Biplot :

To better interpret the PCA result, we utilized biplot to visually represent the data points and variables.

The horizontal x-axis represents projections onto the first principal component (PC1) while the vertical y-axis represents projections onto the second principal component (PC2). The distances between points also align with real differences in country performance.

Data points corresponding to individual countries are plotted based on their scores/coordinates across the two dimensions captured by PC1 and PC2.

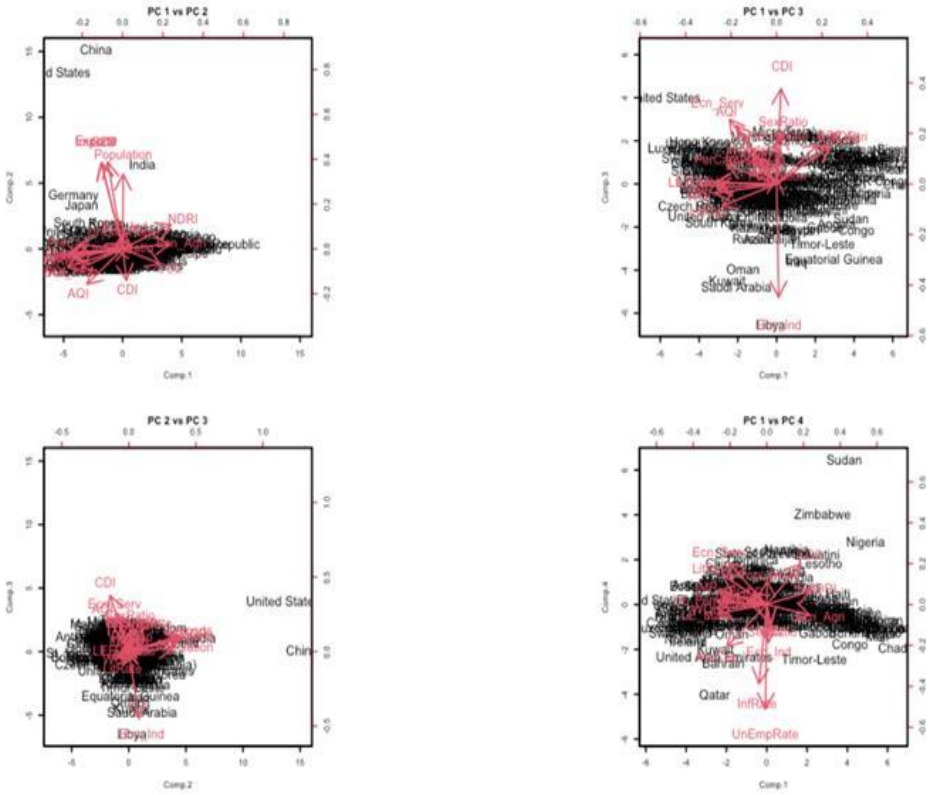

Figure 5.2: PCA - Bi plot

The plot between Comp1 and Comp2 shows that Comp1 is a good measure of the quality of life in a country. It delineates socioeconomic development levels. Countries towards the right of the plot may score lower on development.

PC2 separates along economic productivity/wealth dimensions based on GDP loading. Countries towards the top could are considered more developed/high-income. African countries score above average on principal component 1, while US has the least Comp1. PC2 has high loadings on Population and GDP PerCap, capturing variance explained by economic productivity and income levels. China has the highest Comp2 value. Asian developing countries may fall between as transitioning across both components.

The biplot depicting the relationship between PC2 and PC3 highlights the United States and China as outliers, but in this case, there are no such thing as outliers for this analysis; suggesting that they deviate significantly from the average in terms of Component 2 (Comp2), as they are positioned considerably above the mean. PC3 captures variances related to environmental sustainability.

Another insight from the plots is the strong correlation between Exports and Import which is expected. Import-Export correlation with GDP suggests a connection between a country's international trade and its economic performance. However, a noteworthy insight is the strong correlation between columns CDI and sex ratio. This suggests that Countries that are more culturally diverse (higher CDI scores) tend to also have more gender balanced sex ratios.

Through the analysis of the biplot, we have successfully captured a substantial portion of the economic, environmental, and social factors present in the data. This underscores the efficacy of dimension reduction techniques in revealing the essential patterns within our dataset.

5.4. PCA-3D Plot :

The screenshot shows a 3D scatter plot generated from principal component analysis (PCA). Three principal components (PC1, PC2, PC3) are plotted on the x, y, and z axes respectively.

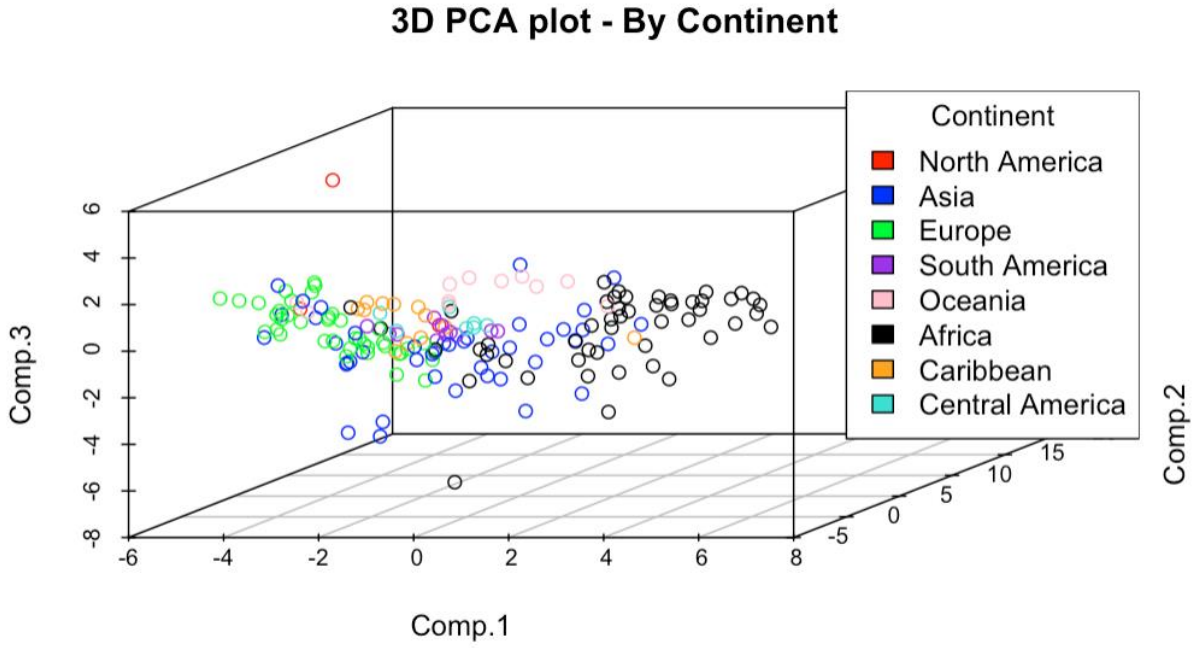

Figure 5.3: Three Dimensions PCA plot

Various data points are visible which represent different countries. Although we cannot see labels for the countries, their positioning provides insight.

Countries cluster together based on how they score across the three dimensions captured by the principal components. PC1 appears to represent socioeconomic development level, with African countries towards the right scoring high.

PC2 deals with economic productivity/income as discussed earlier. Countries towards the right side could be more developed/industrialized.

The loadings implied PC3 links some European and Pacific island nations, so they may group together in that area.

Countries scoring intermediately across the components could be transitioning economically and socially. More dispersed scoring countries represent varied realities.

Through the analysis of the biplot and scatterplot, we successfully captured a substantial portion of the economic, environmental, and social factors present in the data and how variables relate to each other in a reduced dimensional space, providing a framework to interpret real world variations between the continents. This underscores the efficacy of dimension reduction techniques in revealing the essential patterns within our dataset.

5.5. Visually representing PCA on a Map

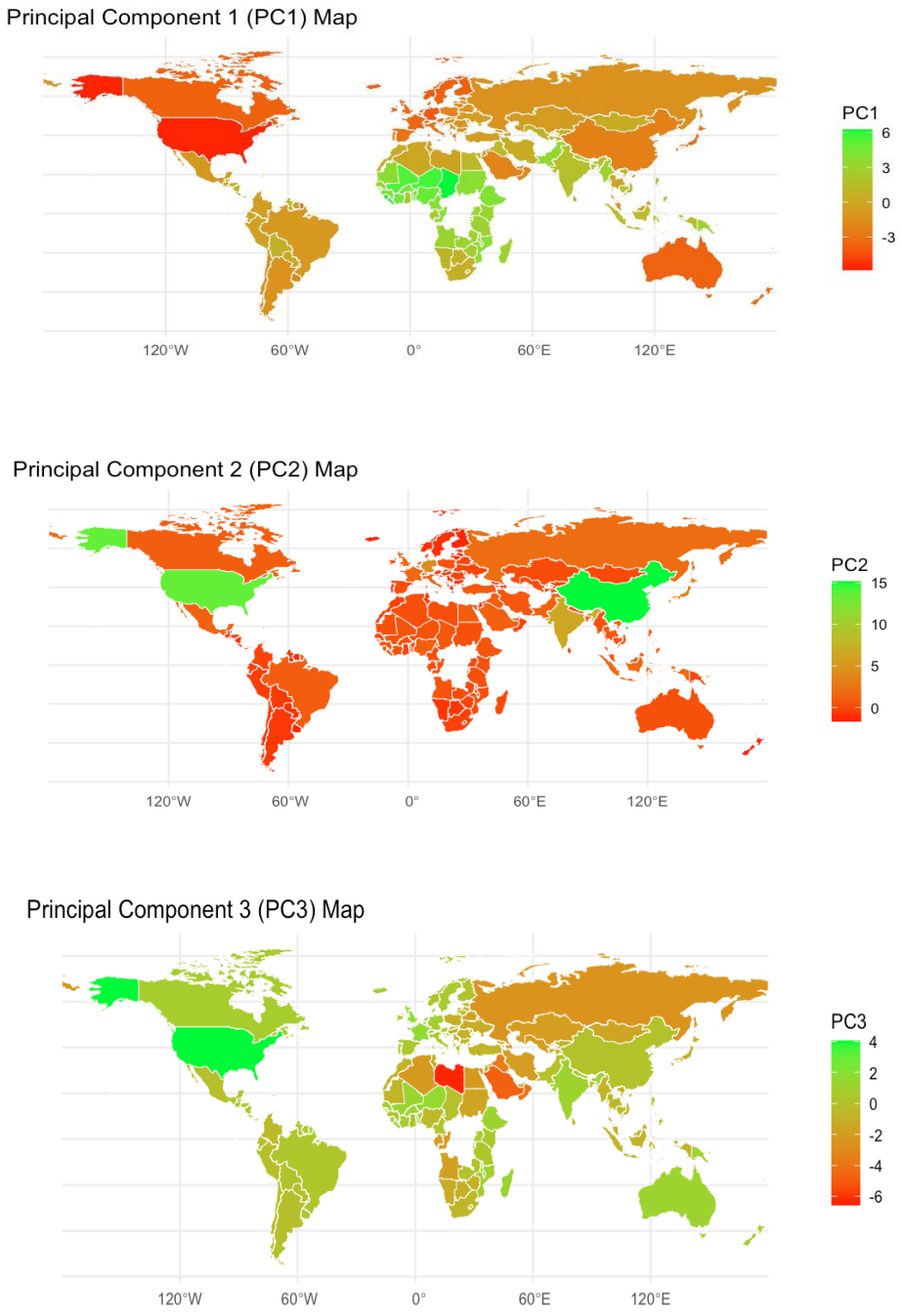

Figure 5.4: Visually Represent principal components on a map.

The map vividly illustrates the geographical distribution of each principal component, evident from the prominent green regions. Specifically, PC1 is notably concentrated in the African continent, while PC2 exhibits a heavy substantial presence in North America and Asia. PC3, on the other hand, extends its influence across various continents, signifying connections among European and Oceanic countries.

5.6. Multi-Dimensional Scaling

We see the coordinate positions of countries. Countries that are closer together in the MDS output are more similar in their socio-economic profiles compared to countries farther apart.

	United States	China	Japan	Germany	India
United States	0.00	3.94	10.51	9.58	10.28
China	3.94	0.00	11.97	11.26	9.54
Japan	10.51	11.97	0.00	1.07	6.06
Germany	9.58	11.26	1.07	0.00	6.30
India	10.28	9.54	6.06	6.30	0.00

Figure 5.5: Distance Matrix of Selected Countries

This low-dimensional visualization allows us to intuitively explore clusters of countries and outliers based on the statistical variables considered. The MDS plot provides insight into cross-country differences and similarities that would be difficult to see from the raw high-dimensional data alone.

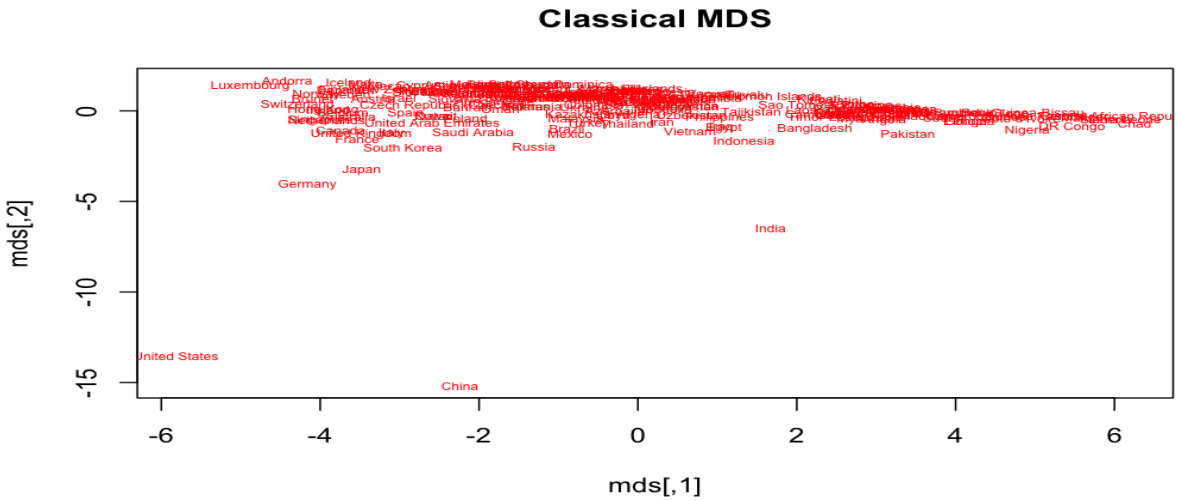

Figure 5.6: Classical MDS plot

To the left of the plot, we see most developed countries, such as the United States, China which are closest to each other and Japan, Germany, and the United Kingdom with similar economic and social conditions. These countries have elevated levels of GDP, per capita income, and life expectancy. We see certain clustering tendencies - Western European nations group together in the top-left area. At the top right of the plot, consists of least developed countries, such as Chad, Niger, and the Democratic Republic of the Congo. These countries are facing the most serious challenges in terms of economic development and social progress.

We performed multidimensional scaling (MDS) on the distance matrix to reduce its dimensionality while preserving pairwise distances between observations. We specified a five-dimensions (k=5) and extracted the eigenvalues.

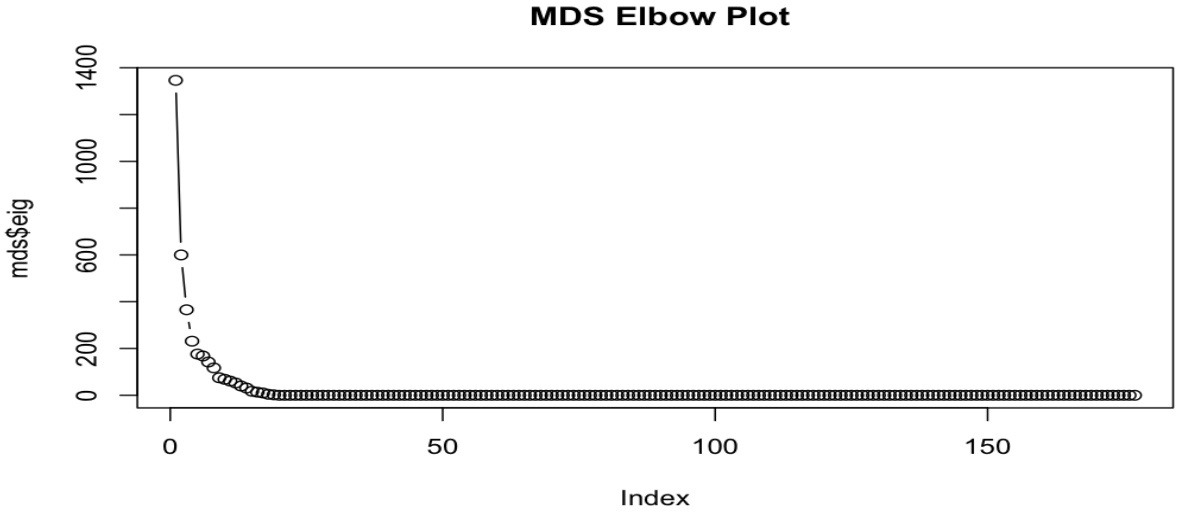

Figure 5.6: MDS Elbow plot

The eigenvalues represent the amount of variance explained by each dimension/component extracted by MDS. To quantify how well the reduced space represents the original distance relationships, we calculated the cumulative proportion of variance explained by the first 5 dimensions.

```
[1] 0.495 0.716 0.850 0.935 1.000
```

The results show that the first dimension alone explains 49.5% of the total variance in the distances. The first two dimensions together explain about 72% of variance. Adding subsequent dimensions increases cumulative variance explained, with the first three dimensions achieving

85% explanation of variance in the data. This indicates that a five-dimensional MDS solution sufficiently captures information in the original distance data, while reducing its complexity for visualization and exploration. Dimensions beyond five contribute minimally to explained variance.

Taking a close look at the MDS plot of the variables, Variables that are closer together correlate more strongly according to the distance metric calculated from the scaled data.

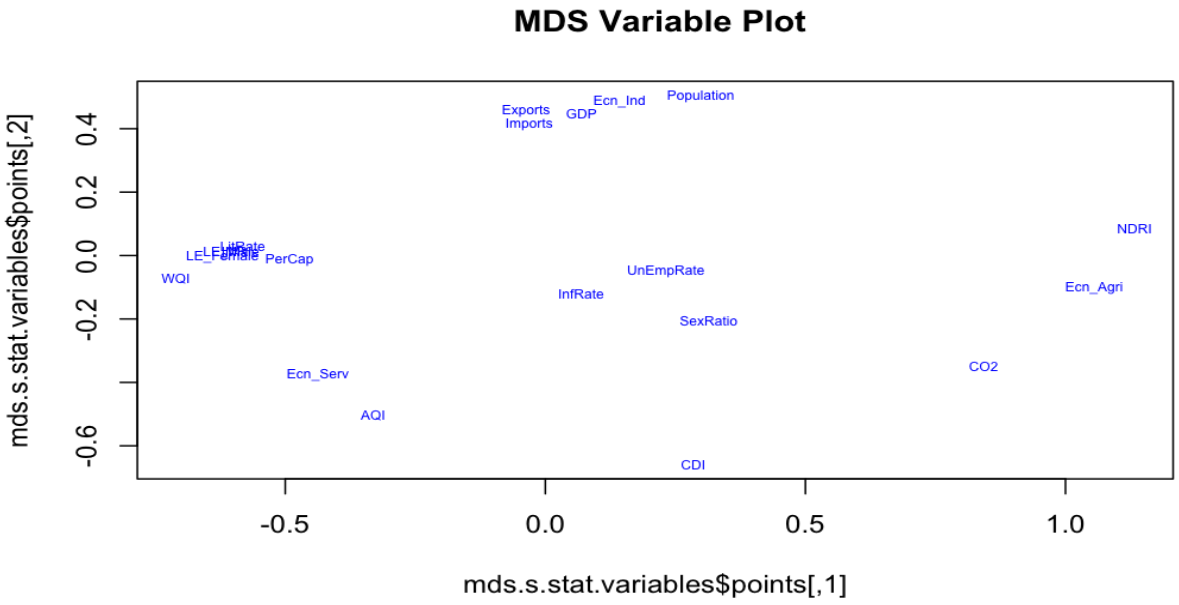

Figure 5.7 : MDS Variable plot

Literacy and Life Exp. cluster very tightly, indicating a close link between education and health outcomes. GDP clusters with import, export, consistent with economic linkages. CO2 emissions lie farther from other variables, suggesting it does not align as closely with socioeconomic factors but closer with environmental variables.

Overall, the MDS mapped out relative proximities between nations based on their holistic profiles, it has also helped visualize similarities in how variables relate across different countries based on their scaled measurements without losing dimensionality like a PCA biplot. Spatial positioning reflects correlation strengths between the various multivariate characteristic of socioeconomic and environmental indicators across this set of global nations.

6. Cluster Analysis

The purpose of cluster analysis is to uncover or discover the groups or clusters of the observations that are homogenous and separated from other groups. Our exploration incorporated Hierarchical clustering with single, complete, and average linkage techniques, as well as K-Means and Model-Based clustering methods. Despite the multifaceted nature of the data, not all clustering results were statistically significant or meaningful. We observed that hierarchical clustering, regardless of the linkage method employed, did not yield robust clusters, suggesting a lack of clear groupings within the data according to these techniques.

Conversely, **K-Means** and **Model-Based** clustering approaches demonstrated more pronounced and significant patterns. These findings suggest differing levels of appropriateness for the clustering methods applied, which we will discuss in detail. The subsequent sections will delve into the specific nuances of the clustering outcomes, and the implications of the significant results obtained from the K-Means and Model-Based clustering analyses.

6.1. Clustering Techniques:

6.1.1. Hierarchical Clustering

Hierarchical clustering is a clustering method where the data is grouped in a series of partitions that may run from a single cluster containing all individuals to n clusters , each containing a single individual.

6.1.2. Hierarchical Clustering - Single Linkage

Single Linkage – Choosing the smallest distance between the groups.

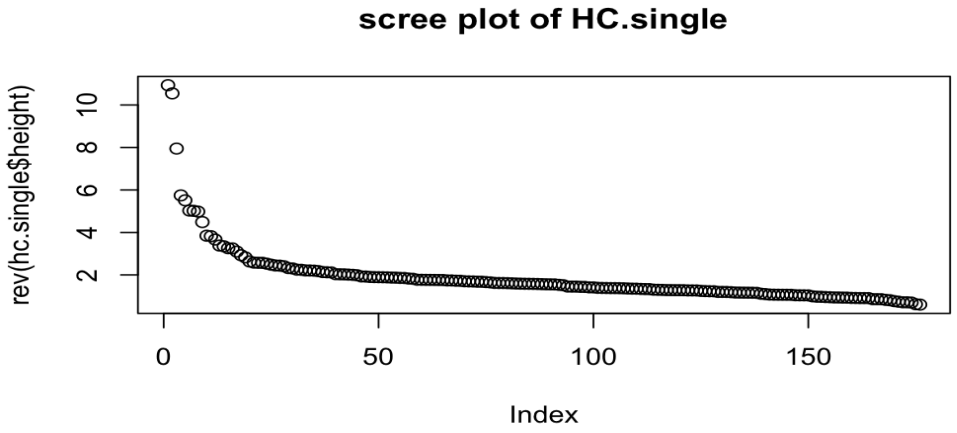

Figure 6.1 : Scree plot for single linkage Hierarchical clustering

Based on the scree plot, set h = 5 as the cut point and we get the below table as clusters outcome.

Number of Each Cluster (Hierarchical Clustering – Single, h = 5)							
Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8
1	1	167	1	3	2	1	1

Table 6.2 : Table of clusters when h = 5 , Single linkage Hierarchical Clustering

The clustering outcome is not good. The clustering outcome shown in the table suggests that most data points (167 out of 177) have been grouped into a single cluster (cluster 3), with the remaining data points being divided among seven other clusters, each containing very few members (only 1 or 2 data points in clusters 1, 2, 4, 7, and 8, and slightly more in clusters 5 & 6).

Changing the number of clusters to interpret the results. Considering the number of clusters as 4 and 5

Number of Each Cluster (Hierarchical Clustering – Single, k = 4)			
Cluster 1	Cluster 2	Cluster 3	Cluster 4
1	1	174	1

Number of Each Cluster (Hierarchical Clustering – Single, k = 5)				
Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
1	1	173	1	1

Table 6.3: Table of clusters when k = 4 , k = 5

Single linkage Hierarchical Clustering

The clustering outcome suggests that the clusters formed by the single linkage method may not be valid or meaningful for this dataset. It implies that the single linkage method did not find a good hierarchical structure in the data.

6.1.3. Hierarchical Clustering - Complete Linkage

Complete Linkage – Choosing the maximum distance between two groups.

Based on the scree plot, set h = 8 as the cut point and we get the below table as clusters outcome.

Number of Each Cluster (Hierarchical Clustering – Complete, h = 8)											
Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10	Cluster 11	Cluster 12
1	1	47	1	97	4	6	2	5	10	2	1

Table 6.4 : Table of clusters when h = 8 , complete linkage Hierarchical Clustering

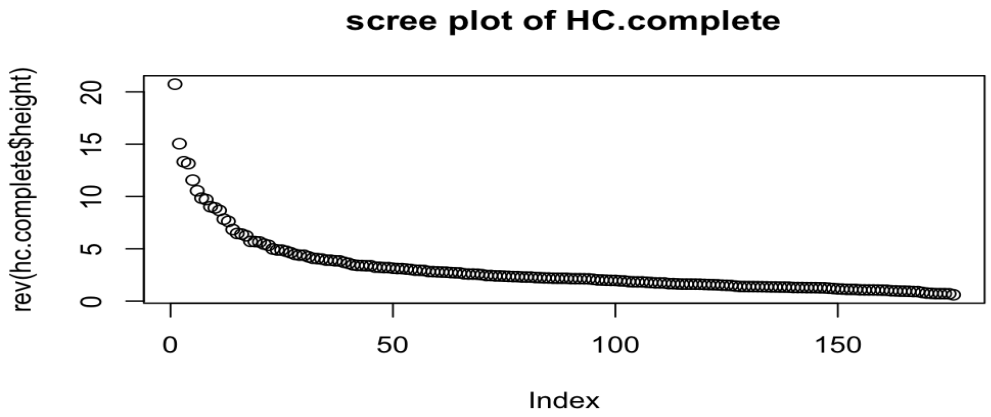

Figure 6.5 : Scree plot for complete linkage Hierarchical clustering

We can observe that the distribution of data points across clusters is quite uneven. Cluster 5 notably contains most data points (97 out of 177), indicating a strong central tendency. Meanwhile, several clusters (Clusters 1, 2, 4 and 12) contain only a single data point, which might suggest that these are outliers or unique cases within the dataset.

Changing the number of clusters to interpret the results. Considering the number of clusters as 3 and 4

Number of Each Cluster (Hierarchical Clustering – Complete, k = 3)		
Cluster 1	Cluster 2	Cluster 3
2	173	2

Number of Each Cluster (Hierarchical Clustering – Complete, k = 4)			
Cluster 1	Cluster 2	Cluster 3	Cluster 4
2	54	119	2

Table 6.6: Table of clusters when k = 3 , k = 4
Complete linkage Hierarchical Clustering

The presence of one or two dominant clusters and several very small clusters suggests that the complete linkage method may not be capturing the natural groupings within the data effectively.

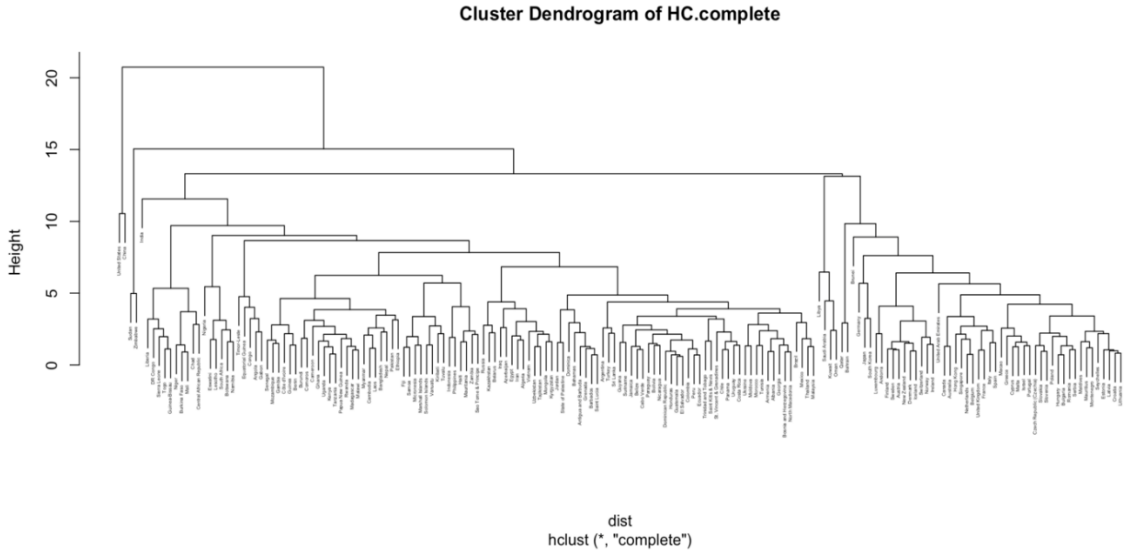

Figure 6.7: Dendrogram - Hierarchical Complete linkage Clustering

6.1.4. Hierarchical Clustering - Average Linkage

Average Linkage – Measuring the average distance between all possible pairs.

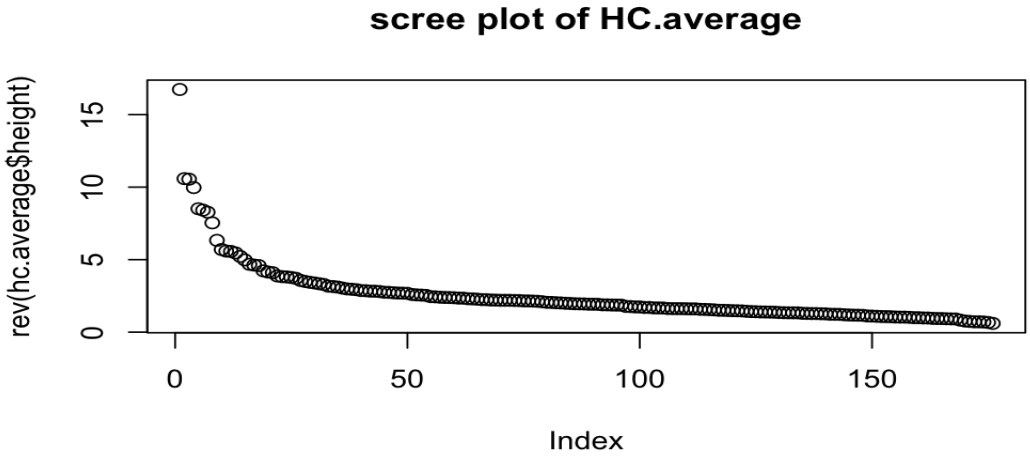

Figure 6.8 : Scree plot for Average linkage Hierarchical clustering

Based on the scree plot, set h = 6 as the cut point and we get the below table as clusters outcome.

Number of Each Cluster (Hierarchical Clustering – Average, h = 6)									
Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10
1	1	121	1	1	4	43	2	2	1

Table 6.9: Table of clusters when h = 6 , Average linkage Hierarchical Clustering

The clustering output for hierarchical clustering using the average linkage method shows a highly imbalanced distribution of data points across the clusters. Notably, Cluster 3 comprises most data points (121 out of 177), while most other clusters contain only one or a few data points.

Changing the number of clusters to interpret the results. Considering the number of clusters as 3 and 4

Number of Each Cluster (Hierarchical Clustering – Average, k = 3)		
Cluster 1	Cluster 2	Cluster 3
2	174	1

Number of Each Cluster (Hierarchical Clustering – Average, k = 3)			
Cluster 1	Cluster 2	Cluster 3	Cluster 4
1	1	174	1

Table 6.10: Table of clusters when k = 3 , k = 4

Average linkage Hierarchical Clustering

Such an imbalanced outcome indicates that the average linkage method may not be effectively distinguishing between distinct groups within the data.

6.2. K - Means Clustering

K – Means clustering is one of the commonly used clustering techniques that tries to find the partition of the n individuals into k groups that minimizes the within – group sum of squares (WGSS) over all variables.

Figure 6.11 : Scree plot for K- Means clustering.

Based on the scree plot, set k = 3 as the cut point and we get the below table as clusters outcome.

Number of Each Cluster (K-Means Clustering, k = 3)		
Cluster 1	Cluster 2	Cluster 3
34	53	90

Table 6.12: Table of clusters when k = 3 , K- Means Clustering

The distribution suggests a balanced partition of the data, with no single cluster dominating. Then we can compute the column means for each cluster and assign labels to each one.

Column Means for each Cluster - kmeans

	Population	LE_Male	LE_Female	UnEmpRate	SexRatio	IMR	LitRate	CDI	GDP	PerCap	InfRate	Exports	Imports	Ecn_Agri	Ecn_Ind	Ecn_Serv	NDRI	CO2	AQI	WQI
Cluster 1	0.182	1.210	1.117	0.457	0.360	0.856	0.662	0.169	0.676	1.67	0.254	0.998	0.978	-0.83	-0.227	0.905	-0.984	-1.066	0.636	1.089
Cluster 2	0.124	-1.225	-1.269	0.205	0.147	-1.217	-1.201	0.271	-0.175	-0.68	-0.197	-0.285	-0.275	1.04	-0.022	-0.873	1.155	0.756	-0.769	-1.280
Cluster 3	-0.142	0.264	0.326	-0.294	-0.223	0.393	0.457	-0.223	-0.152	-0.23	0.020	-0.209	-0.208	-0.30	0.099	0.172	-0.309	-0.042	0.213	0.342

Table 6.13: ColumnMeans, K- Means Clustering

- Cluster 1 appears to have relatively higher Life Expectancy (LE) for both males and females, higher Infant Mortality Rate (IMR), Literacy Rate (LitRate), higher GDP and GDP per capita, et cetera. This could suggest a cluster of countries/regions with better

socio-economic status and environmental indicators, so a possible label could be "Higher Socio-economic Status".

- Cluster 2 shows negative values for Male and Female LE, and low economic indicators, such as GDP per capita, Exports and Imports. This might represent countries/regions with poor socio-economic conditions and lower economic indicators, so a label like "Lower Socio-economic Status" might be appropriate.
- Cluster 3 has a mix of positive and negative values across different indicators, all the indicators are not very big, this cluster might represent "Stagnant or Declining Economies".

6.2.1. Interpreting K-Means based on PCA.

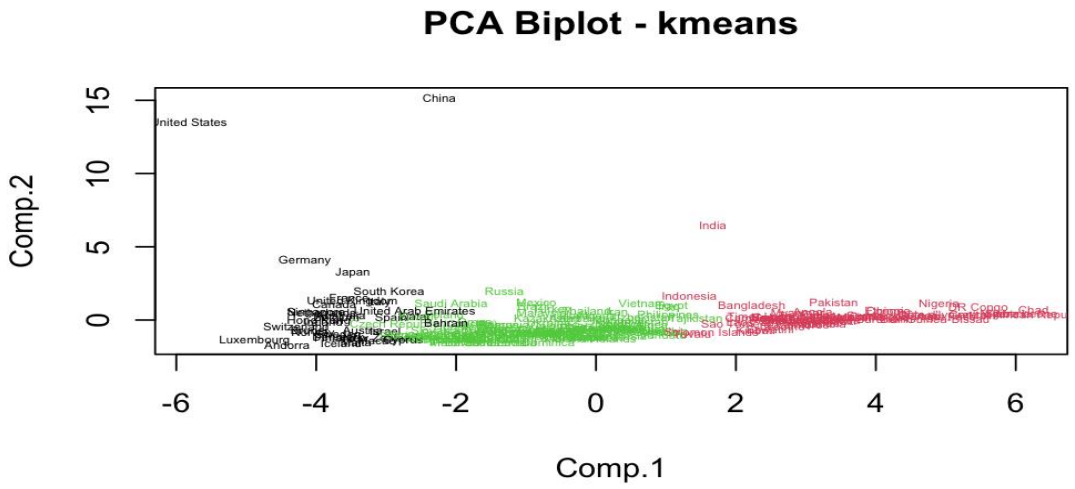

Figure 6.14 : Biplot for PCA , K- Means Clustering

Based on the PCA Biplot – K-Means, we gain insight that countries like US, China, Germany, and Japan are marked by high economic productivity. Conversely, countries including India, Pakistan, Nigeria are highlighted for their improved quality of life. Meanwhile, countries like Russia, Mexico, Vietnam are characterized by their stagnant economics.

We could also make a map to have a more intuitive view.

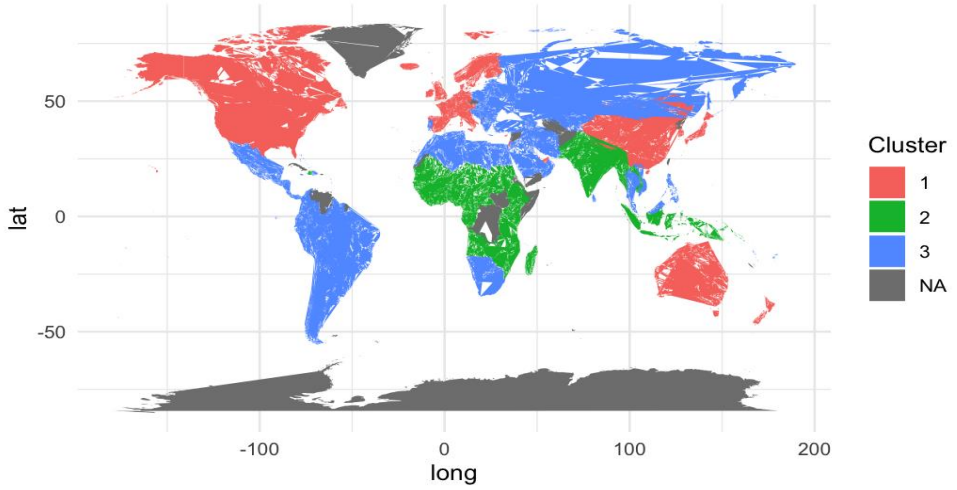

Figure 6.15 : Map showing the K-Means clustering.

6.3. Model-Based clustering

Model – Based clustering technique will group the clusters or groups assuming that the clusters are forming from a normal distribution.

In Model – Based clustering to determine the number of clusters we measure the BIC (Bayes information criteria) which is based on the trade-off between the simplicity and the quality of the model.

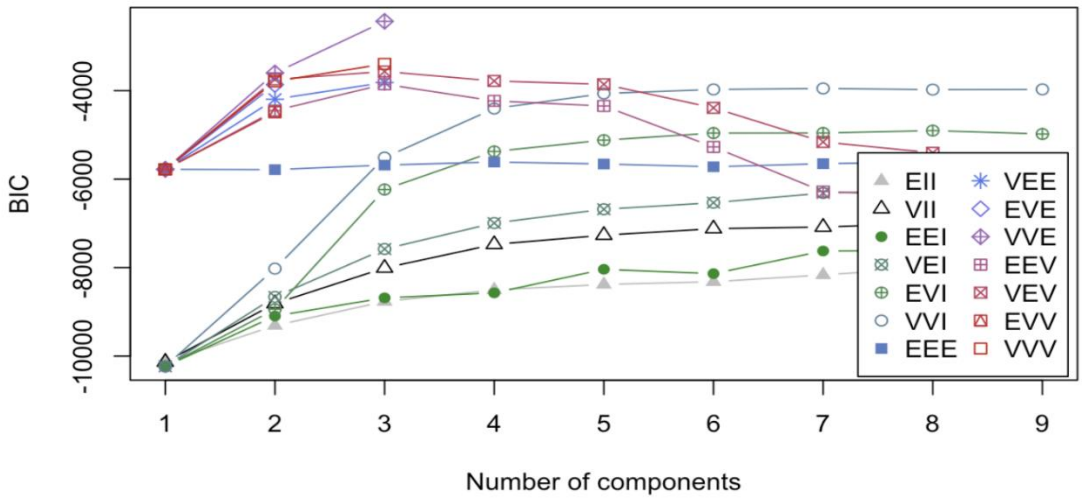

Figure 6.16: BIC plot based on Model-Based Clustering

The summary of the BIC plot can also be used to estimate the number of clusters, the below table shows the summary of the BIC.

```
Best BIC values:
             VVE,3      VVV,3     VEV,3
BIC      -2431.773 -3394.8562 -3566.554
BIC diff     0.000  -963.0829 -1134.781
```

Figure 6.17: Summary of the BIC

From the above BIC plot and the BIC summary , we can infer that the number of clusters suitable for model- based clustering are 3.

Uncertainty values which refer to a measure that quantifies the confidence with which the datapoint is assigned to a certain cluster. The Uncertainty of the cluster values can also be evaluated by using the results of Mclust. Below is the table for Uncertainty .

	[,1]	[,2]	[,3]
Barbados	"Barbados"	"3"	"0.188911390359512"
Belgium	"Belgium"	"2"	"0.0857266523386688"
Bosnia and Herzegovina	"Bosnia and Herzegovina"	"2"	"0.0440745721656712"
Uruguay	"Uruguay"	"2"	"0.012176294755217"
Dominican Republic	"Dominican Republic"	"3"	"0.00529394946799244"
Costa Rica	"Costa Rica"	"3"	"0.00252606028309099"
Azerbaijan	"Azerbaijan"	"2"	"0.00229722938966404"
Netherlands	"Netherlands"	"1"	"0.0006478346269716"
Montenegro	"Montenegro"	"3"	"0.000621657986781043"
Bahamas	"Bahamas"	"2"	"0.000565065187716818"

Figure 6.18: Uncertainty of the given observation

From the above uncertainty values , we can infer that the countries Barbados from cluster 3 , Belgium , Bosnia and Herzegovina from cluster 2 are the top three countries with maximum uncertainty.

The clustering outcome is as follows.

Number of Each Cluster (Model-Based Clustering, k = 3)		
Cluster 1	Cluster 2	Cluster 3
26	64	87

Table 6.19: ColumnMeans, K- Means Clustering

The distribution suggests a balanced partition of the data, with no single cluster dominating. Then we can compute the column means for each cluster and assign labels to each one.

Column Means for each Cluster - ModelBased

	Population	LE_Male	LE_Female	UnEmpRate	SexRatio	IMR	LitRate	CDI	GDP	PerCap	InfRate	Exports	Imports	Ecn_Agri	Ecn_Ind	Ecn_Serv	NDRI	CO2	AQI	WQI
Cluster 1	0.837	0.411	0.340	-0.196	-0.283	0.310	0.287	-0.647	0.934	0.526	-0.640	1.131	1.121	-0.479	0.384	0.082	-0.286	-1.250	-0.612	0.309
Cluster 2	-0.081	0.673	0.716	0.174	0.004	0.637	0.542	-0.113	-0.080	0.572	0.130	0.004	-0.008	-0.569	0.055	0.440	-0.585	-0.330	0.419	0.692
Cluster 3	-0.192	-0.620	-0.630	-0.069	0.082	-0.563	-0.486	0.278	-0.221	-0.580	0.096	-0.343	-0.331	0.563	-0.156	-0.349	0.517	0.619	-0.125	-0.603

Table 6.20: ColumnMeans, Model based Clustering.

- Cluster 1 has high positive means in Population, Life Expectancy (both male and female), GDP, and GDP Per Capita, but a negative mean for CDI and a significant negative mean for CO2 emissions. This cluster might represent countries with strong economic performance but lower environmental impact. This cluster can be called "Economic Performers."
- Cluster 2 with positive means in Life Expectancy, Literacy Rate, and GDP Per Capita, coupled with moderate values in Economic, Social, and Environmental, factors this cluster seems to represent countries with balanced socio-economic development and a healthier quality of life. This cluster can be called "Balanced Development".
- Cluster 3 has negative meanings for Life Expectancy, Literacy Rate, and GDP Per Capita, but higher values for Agriculture. It also has a negative mean for the Water Quality Index. These countries might be in a developing stage with challenges in human development and environmental quality. This cluster can be called "Developing with Challenges".

6.3.2. Interpreting Model-Based Clusters based on PCA.

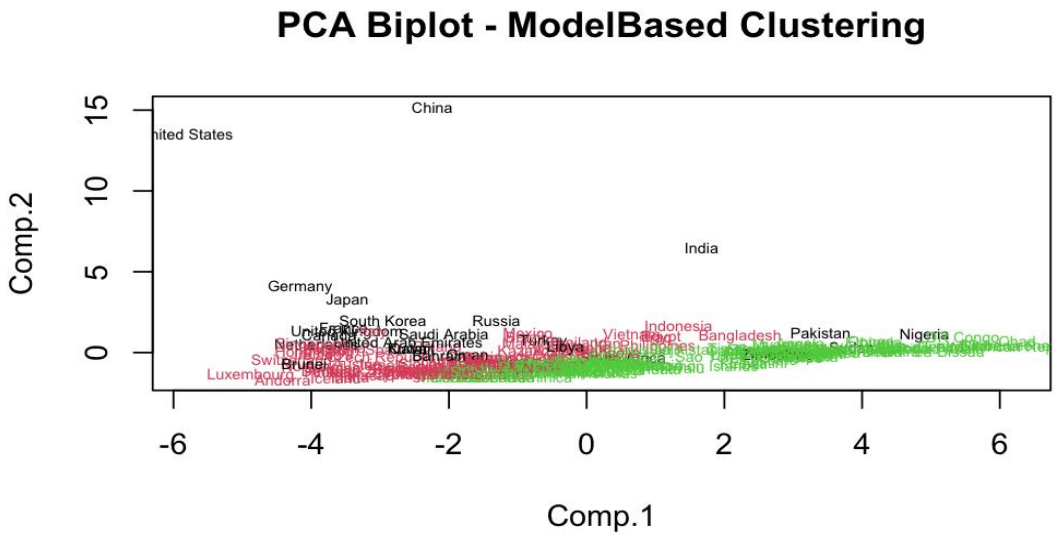

Figure 6.21 : Biplot for PCA , Model Based Clustering

Nations such as the United States, China, and Germany exemplify robust economic performance, while countries such as Indonesia and Vietnam demonstrate balanced socio-economic growth. Meanwhile, countries such as Nepal and the DR Congo are in the developmental phase, facing various socio-economic challenges. We can also make a map.

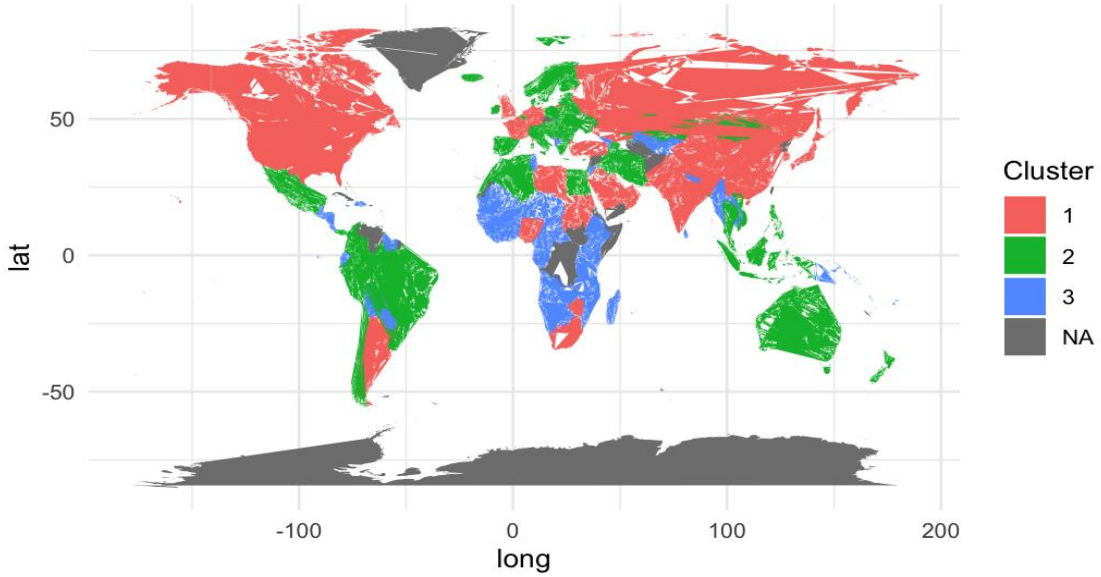

Figure 6.22 : Map showing the model-based clustering.

However, the PCA Biplot for Model-Based Clustering does not appear to be very satisfactory, as it exhibits overlapping components.

6.4. Evaluating the clustering outcomes

Since we do not have a true cluster, it is hard to use chisq.test to check which model is better. We could use Silhouette Coefficient to examine which one is better.

Silhouette Coefficient[6]

- **Formula**: The silhouette coefficient for a single sample is (b-a)/max(a,b) where *a* is the mean distance to the other instances in the same cluster, and *b* is the mean nearest-cluster distance, i.e., the distance to the instances of the next closest cluster.
- **Interpretation**: Values range from -1 to 1. A high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters.

6.4.1. K-Means Clustering - Results:

Average Silhouette Widths: The average silhouette widths for the three clusters are approximately 0.3383, 0.2396, and 0.0566, respectively.

Individual Silhouette Widths: The individual silhouette widths for each data point within the clusters range from -0.154 to 0.491, with a mean value of approximately 0.255. In the K-Means clustering results, the average silhouette widths for all three clusters are positive, indicating that the data points within each cluster are well-matched to their own cluster compared to neighboring clusters.

```
Silhouette of 177 units in 3 clusters from silhouette.default(x = km$cluster, dist =
dist(s.statistics)) :
 Cluster sizes and average silhouette widths:
    90     53     34
0.3383 0.2396 0.0566
Individual silhouette widths:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -0.154   0.144   0.295   0.255   0.380   0.491
```

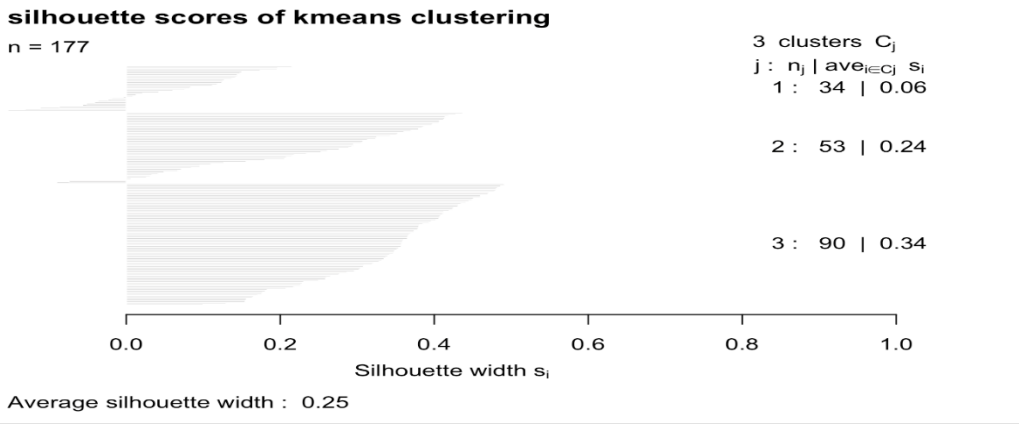

Figure 6.23: Represents the silhouette widths of K-Means Clustering

6.4.2. Model-Based Clustering - Results:

Average Silhouette Widths: The average silhouette widths for the three clusters are approximately -0.276, 0.302, and 0.114, respectively.

Individual Silhouette Widths: The individual silhouette widths for each data point within the clusters range from -0.464 to 0.479, with a mean value of approximately 0.125.

In the Model-Based clustering results, the average silhouette width for one of the clusters is negative (-0.276), which suggests that the data points within that cluster are more like data points in neighboring clusters than to their own cluster. This indicates that the clustering quality for this method is not as good as in K-Means.

Overall, based on the silhouette scores, the K-Means clustering appears to perform better than Model-Based clustering in this case. K-Means shows more distinct and well-separated clusters, as indicated by the positive average silhouette widths for all clusters and the higher mean silhouette width.

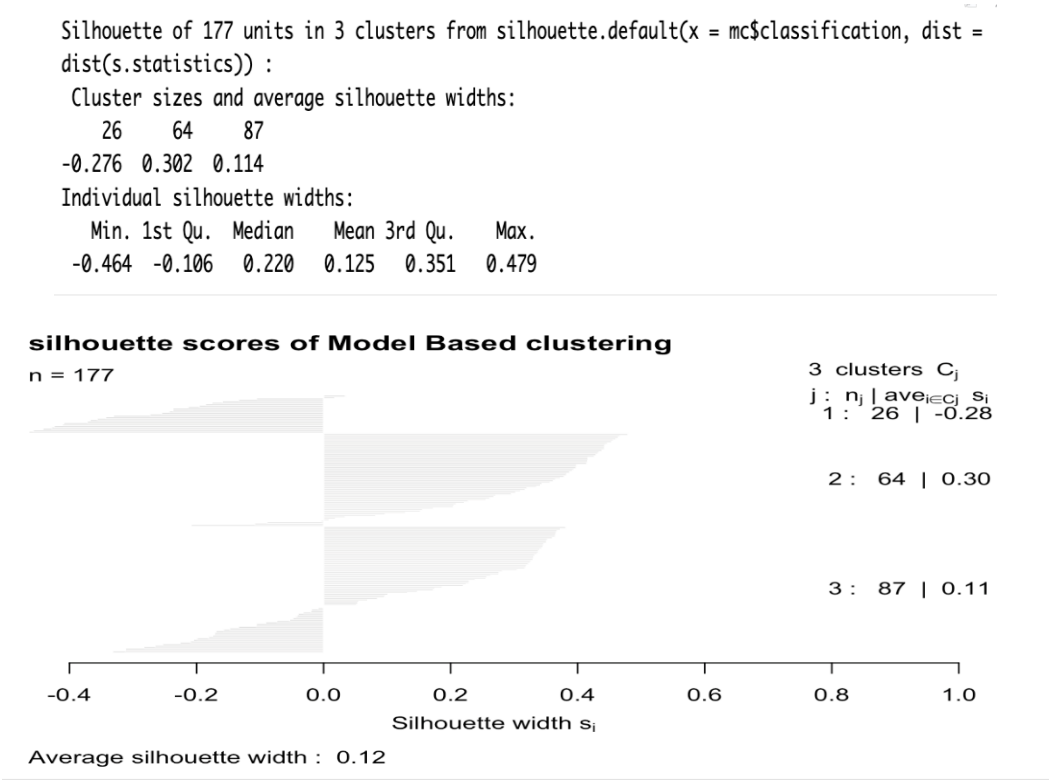

Figure 6.24: Represents the silhouette widths of K-Means Clustering

7. Confirmatory Factor Analysis

7.1. Factor Analysis.

Factor Analysis uncovers the relationship between the latent variables that are assumed to the manifest variables. Variables that are hidden and quantifies the variables that are known (or present in the dataset) can be termed as the **Latent Variables** or the factors . And the known variables are called the **Manifest Variables.**

7.2. UN Statistics Data Set:

In the UN Statistics Dataset, we have 24 columns these are the manifest variables which are used to interpret the latent variables (factors)

Manifest Variables	Country, Population, LE_Male, LE_Female, UnEmpRate, SexRatio, IMR, LitRate, CDI, GDP, PerCap, InfRate, Exports, Imports ,Ecn_Agri, Ecn_Ind, Ecn_Serv, NDRI, NDRI_Cat, CO2, AQI, AQI_Cat, WQI, WQI_Cat
Latent Variables	Economic Indicators , Environmental Indicators , Social Indictors

Table 7.1 : Table showing the Manifest and Latent Variables

7.3. Confirmatory Factor Analysis:

The Confirmatory Factor Analysis model is used to identify which manifest variables are related to a particular factor when the other manifest variables are restricted to having zero loadings on the other factors.

7.3.1. Requirements for Confirmatory Factor Analysis:

Confirmatory Factor Analysis model can be developed based on the theoretical considerations from a given dataset or can be based on the factors that are the results of the Exploratory Factor Analysis (EFA) model.

In our project, we developed factors using Exploratory Factor Analysis that are further utilized for specifying model for Confirmatory Factor Analysis. Firstly , considered taking all the numeric manifest variables to perform the Exploratory factor analysis to develop a theory for our factor analysis.

In exploratory factor analysis (EFA), factor loadings indicate the strength and direction of the relationship between observed variables (indicators) and latent factors. A negative factor loading means that there is a negative correlation between the observed variable and the corresponding factor. In other words, as the value of the observed variable increases, the expected value of the latent factor decreases, and vice versa.

In confirmatory factor analysis (CFA), factor loadings are typically constrained to be positive because they represent the standardized regression coefficients between observed variables (indicators) and latent factors. These loadings indicate the strength and direction of the relationship between the observed variables and the latent constructs.

In CFA, the idea is to test a hypothesized model where you specify the expected relationships between observed variables and latent factors based on theory or prior research. Allowing for negative factor loadings in a CFA model might complicate the interpretation of the model, as negative loadings are generally not consistent with the typical understanding of factor loadings in CFA.

7.3.2. Results for the Exploratory factor analysis assuming 3 factors.

```
Call:
factanal(factors = 3, covmat = statistics.corr, n.obs = 177)

Uniquenesses:
Population    LE_Male  LE_Female  UnEmpRate   SexRatio        IMR    LitRate        CDI        GDP     PerCap    InfRate
     0.622      0.111      0.069      0.979      0.977      0.160      0.337      0.855      0.100      0.501      0.985
   Exports    Imports   Ecn_Agri    Ecn_Ind   Ecn_Serv       NDRI        CO2        AQI        WQI
     0.038      0.005      0.296      0.005      0.216      0.238      0.606      0.544      0.095

Loadings:
           Factor1 Factor2 Factor3
Population -0.109   0.601
LE_Male     0.938
LE_Female   0.960
UnEmpRate           0.121
SexRatio                    0.126
IMR         0.914
LitRate     0.807          -0.102
CDI                -0.106   0.359
GDP                 0.944
PerCap      0.672   0.216
InfRate     0.112
Exports     0.196   0.958
Imports     0.193   0.978
Ecn_Agri   -0.760           0.352
Ecn_Ind                    -0.996
Ecn_Serv    0.691           0.550
NDRI       -0.840   0.130   0.197
CO2        -0.566  -0.172   0.211
AQI         0.608  -0.108   0.273
WQI         0.951

                Factor1 Factor2 Factor3
SS loadings       7.216   3.290   1.757
Proportion Var    0.361   0.164   0.088
Cumulative Var    0.361   0.525   0.613

Test of the hypothesis that 3 factors are sufficient.
The chi square statistic is 2048.41 on 133 degrees of freedom.
The p-value is 0
```

Figure 7.2 : Exploratory Factor analysis assuming 3 factors

From the results of the exploratory factor analysis (EFA) , we can see that the factor loadings for a few of the variables are quite low, that is some are less than 0.2 and some are even negative. This means that the relationship between the factors (latent variables and the manifest variables is quite low. And a negative sign in the factor loadings indicates a negative correlation we cannot be suitable for specifying a confirmatory factor analysis (CFA) model.

So, we exclude the negative loading variables and the variables with loadings less than 0.4 and all the negative coefficients and prepare our data for theoretical model used for the confirmatory factor analysis. Let's find the correlation matrix between the variables and observe the values.

	Population	LE_Male	LE_Female	IMR	LitRate	GDP	PerCap	Exports	Imports	Ecn_Serv	AQI	WQI
Population	1.00	0.00	-0.02	0.00	-0.02	0.58	-0.05	0.62	0.56	-0.08	-0.23	-0.18
LE_Male	0.00	1.00	0.96	0.87	0.72	0.15	0.66	0.28	0.28	0.60	0.51	0.87
LE_Female	-0.02	0.96	1.00	0.92	0.76	0.15	0.63	0.27	0.27	0.66	0.55	0.89
IMR	0.00	0.87	0.92	1.00	0.80	0.15	0.52	0.24	0.24	0.64	0.54	0.84
LitRate	-0.02	0.72	0.76	0.80	1.00	0.13	0.45	0.20	0.19	0.55	0.49	0.78
GDP	0.58	0.15	0.15	0.15	0.13	1.00	0.22	0.91	0.94	0.10	-0.06	0.07
PerCap	-0.05	0.66	0.63	0.52	0.45	0.22	1.00	0.34	0.34	0.49	0.46	0.69
Exports	0.62	0.28	0.27	0.24	0.20	0.91	0.34	1.00	0.98	0.15	0.00	0.18
Imports	0.56	0.28	0.27	0.24	0.19	0.94	0.34	0.98	1.00	0.17	0.00	0.18
Ecn_Serv	-0.08	0.60	0.66	0.64	0.55	0.10	0.49	0.15	0.17	1.00	0.62	0.65
AQI	-0.23	0.51	0.55	0.54	0.49	-0.06	0.46	0.00	0.00	0.62	1.00	0.63
WQI	-0.18	0.87	0.89	0.84	0.78	0.07	0.69	0.18	0.18	0.65	0.63	1.00

Figure 7.3 : Correlation between the variables after suitably excluding some of the variables.

The observed variables that load onto a latent factor are expected to be related to that factor. The presence of variables with zero correlation might raise questions about the conceptual coherence of including those variables in the same latent factor. If two variables are essentially unrelated, it might be more appropriate to consider them as indicators of separate latent factors or to reconsider their inclusion in the model.

However, it's worth noting that "zero correlation" does not necessarily mean "unrelated." Sometimes, variables might be conceptually related but not correlated due to measurement error or other factors. In such cases, it's important to carefully consider the theoretical and practical implications of including or excluding those variables in the CFA model.

```
Population    LE_Male  LE_Female       IMR    LitRate       GDP     PerCap    Exports    Imports   Ecn_Serv        AQI
      0.64       0.07       0.01      0.14       0.40      0.10       0.55       0.04       0.00       0.56       0.66
       WQI
      0.18

Loadings:
           Factor1 Factor2
Population          0.595
LE_Male     0.961
LE_Female   0.988
IMR         0.923
LitRate     0.772
GDP                 0.947
PerCap      0.626
Exports             0.963
Imports             0.981
Ecn_Serv    0.663
AQI         0.577
WQI         0.908

                Factor1 Factor2
SS loadings       5.414   3.238
Proportion Var    0.451   0.270
Cumulative Var    0.451   0.721
```

Figure 7.4 : Results of the Exploratory factor analysis assuming 2 factors

From the newly created data set , we assume two factors and perform exploratory factor analysis. And get the above results. From the above results we can see that all the loadings uniquely identify and also satisfies our criteria for having non-negative and loadings that are greater than 0.5. We can also see that the uniqueness of the variable Imports is zero which means that there might be a multicollinearity between variables which can complicate the interpretation of results. So, let's exclude imports variable and after repetitive iterations , and to reduce the multicollinearity we came up with a specific models that can test and confirm that the assumed model provides adequate for the correlations among the manifest variables.

After considering all the above factors , the finalized EFA model or the theoretical model to perform confirmatory factor analysis is as follows.

```
Loadings:
           Factor1 Factor2
GDP                 0.971
Imports             0.949
Population          0.602
LE_Male     0.891
IMR         0.974
LitRate     0.815

                Factor1 Factor2
SS loadings        2.46    2.22
Proportion Var     0.41    0.37
Cumulative Var     0.41    0.78
       GDP    Imports Population    LE_Male        IMR    LitRate
      0.05       0.06       0.64       0.20       0.05       0.33
```

Figure 7.5 : Finalized Exploratory Factor analysis results , assuming 2 factors

7.4.CFA Model :

7.4.1. Model1: Suggested model contains "GDP", "Imports", "Population", "LE_Male", "IMR", "LitRate" as manifest variables and health and economy as the latent factors.

Using the results of the Exploratory Factor Analysis , we created a specify Model assuming the latent factors health (living or social) conditions of the countries depends on the manifest variables like the life Expectancy of the male (female) , the infant mortality rate and the literacy rate. The Economic indicators may vary depending on the manifest variables like the Gross Domestic Product , the Imports (Exports) and the population.

	Estimate <dbl>	Std Error <dbl>	z value <dbl>	Pr(>|z|) <dbl>	<chr>
0.891	0.889	0.060	14.9	3.7e-50	LE_Male <--- Health
0.974	0.980	0.056	17.7	8.6e-70	IMR <--- Health
0.815	0.814	0.063	13.0	1.2e-38	LitRate <--- Health
0.971	0.980	0.056	17.7	9.2e-70	GDP <--- Economy
0.949	0.961	0.057	17.0	8.1e-65	Imports <--- Economy
0.602	0.591	0.069	8.5	1.5e-17	Population <--- Economy
rho	0.189	0.075	2.5	1.2e-02	Economy <--> Health
0.20	0.210	0.030	7.0	2.4e-12	LE_Male <--> LE_Male
0.05	0.039	0.022	1.7	8.1e-02	IMR <--> IMR
0.33	0.338	0.040	8.4	3.9e-17	LitRate <--> LitRate
0.06	0.077	0.024	3.2	1.3e-03	Imports <--> Imports
0.64	0.651	0.070	9.2	2.5e-20	Population <--> Population

Figure 7.6: Model 1- Estimated parameters for the CFA model

From the above results ,the correlation between Health and Economy is known as the disattenuated correlation . The disattenuated correlation between Health and Economic indicators of the UN statistics is 0.189.

Let's calculate the 95% confidence interval for the disattenuated correlation.

Confidence Interval – Model 1 :

$$\text{Confidence Interval} = \text{Estimate} \pm \text{CriticialValue} * \text{Std Error}$$

After computing the Confidence Interval for the disattenuated correlation for 95 % confidence level is [0.042, 0.34]

Similarly, we can Estimate the confidence intervals for all the manifest variables.

Even though we do not have the p-value > 0.05 , we cannot rely on p-value, so we move forward with our analysis to estimate the root mean square error.

The goodness of fit of a model depends on the discrepancy between the actual correlation matrix and the estimated correlation matrix.

Restricted cor matrix

	GDP	Imports	Population	LE_Male	IMR	LitRate
GDP	1.00	0.94	0.58	0.16	0.18	0.15
Imports	0.94	1.00	0.57	0.16	0.18	0.15
Population	0.58	0.57	1.00	0.10	0.11	0.09
LE_Male	0.16	0.16	0.10	1.00	0.87	0.72
IMR	0.18	0.18	0.11	0.87	1.00	0.80
LitRate	0.15	0.15	0.09	0.72	0.80	1.00

Figure 7.7 : Restricted or the estimated correlation matrix of sem - Model 1

Non – restricted or the original correlation matrix

	GDP	Imports	Population	LE_Male	IMR	LitRate
GDP	1.00	0.94	0.58	0.15	0.15	0.13
Imports	0.94	1.00	0.56	0.28	0.24	0.19
Population	0.58	0.56	1.00	0.00	0.00	-0.02
LE_Male	0.15	0.28	0.00	1.00	0.87	0.72
IMR	0.15	0.24	0.00	0.87	1.00	0.80
LitRate	0.13	0.19	-0.02	0.72	0.80	1.00

Figure 7.8 :Non-Restricted or the actual correlation matrix of sem - Model 1

7.4.2. Root Mean square error – Model 1 :

The root mean square error (RMSE) measures the average difference between the predicted correlation matrix and the actual correlation matrix.

The root mean square error value for our estimated and original matrix is 0.055 which is less than 0.1 , so our model is acceptable.

Now we estimate other indexes that are used to test the goodness fit of the model. We use GFI (Goodness- of fit index , AGFI (Adjusted Goodness of fit index) , SRMR (Square root Mean Residual) – which is the root mean square error computed above.

```
Model Chisquare =  35   Df =  9 Pr(>Chisq) = 6e-05
Goodness-of-fit index =  0.94
Adjusted goodness-of-fit index =  0.86
SRMR =  0.051

Normalized Residuals
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -1.49   -0.32    0.00   -0.13    0.01    1.52

R-square for Endogenous Variables
  LE_Male        IMR    LitRate        GDP    Imports Population
     0.79       0.96       0.66       0.96       0.92       0.35

Parameter Estimates

Iterations =  39
```

Figure :7.9 Parameter Estimations for goodness of fit - Model1

From the above table , we can say that the GFI is 0.94 , AGFI is 0.86 and SRMR is 0.051 . Any model with GFI greater than 0.9 , AGFI greater than 0.8 and SRMR less than 0.1 can be considered a good model. So, we can say that our model is good and acceptable.

To visually represent our model , we use path diagram which explains the correlation between manifest variables, the correlation between the factors and the variances between the manifest variables (uniqueness) and the covariance between the factors.

7.4.3. Path Diagram for our specified CFA model 1

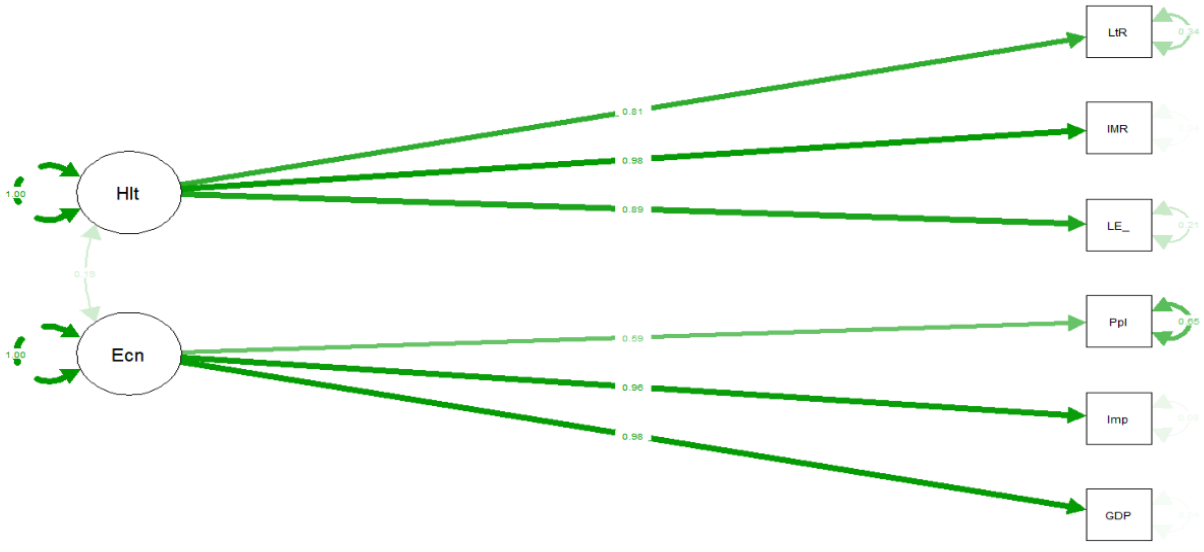

Figure 7.10: Path Diagram for CFA model – Model1

7.4.4. Model 2:

Suggested model contains "GDP", "Exports", "Population", "LE_Female","IMR", "LitRate" as manifest variables and health and economy as the latent factors.

Using the results of the Exploratory Factor Analysis , we created a specify Model assuming the latent factors health (living or social) conditions of the countries depends on the manifest variables like the life Expectancy of the male (female) , the infant mortality rate and the literacy rate. The Economic indicators may vary depending on the manifest variables like the Gross Domestic Product , the Imports (Exports) and the population.

	Estimate <dbl>	Std Error <dbl>	z value <dbl>	Pr(>|z|) <dbl>	<chr>
0.942	0.941	0.057	16.43	1.1e-60	LE_Female <--- Health
0.975	0.981	0.055	17.70	4.6e-70	IMR <--- Health
0.814	0.814	0.062	13.05	6.0e-39	LitRate <--- Health
0.914 NA	0.902	0.061	14.80	1.6e-49	GDP <--- Economy
0.977	1.011	0.057	17.83	4.0e-71	Exports <--- Economy
0.642	0.608	0.069	8.79	1.5e-18	Population <--- Economy
rho	0.268	0.070	3.81	1.4e-04	Economy <--> Health
0.11	0.115	0.023	4.93	8.3e-07	LE_Female <--> LE_Female
0.05	0.039	0.022	1.76	7.8e-02	IMR <--> IMR
0.34	0.338	0.039	8.66	4.7e-18	LitRate <--> LitRate
0.16	0.186	0.039	4.76	1.9e-06	GDP <--> GDP
0.00	-0.022	0.042	-0.53	6.0e-01	Exports <--> Exports

Figure 7.11: Model 2- Estimated parameters for the CFA model

From the above results ,the correlation between Health and Economy is known as the disattenuated correlation. The disattenuated correlation between Health and Economic indicators of the UN statistics is 0.268.

Let's calculate the 95% confidence interval for the disattenuated correlation.

Confidence Interval – Model 2 :

$$\text{Confidence Interval} = \text{Estimate} \pm \text{CriticialValue} * \text{Std Error}$$

After computing the Confidence Interval for the disattenuated correlation for 95 % confidence level is [0.14, 0.41]

Similarly, we can Estimate the confidence intervals for all the manifest variables.

Even though we do not have the p-value > 0.05 , we cannot rely on p-value, so we move forward with our analysis to estimate the root mean square error.

The goodness of fit of a model depends on the discrepancy between the actual correlation matrix and the estimated correlation matrix.

Restricted cor matrix

```
            GDP Exports Population LE_Female  IMR LitRate
GDP        1.00    0.91       0.55      0.23 0.24    0.20
Exports    0.91    1.00       0.61      0.26 0.27    0.22
Population 0.55    0.61       1.00      0.15 0.16    0.13
LE_Female  0.23    0.26       0.15      1.00 0.92    0.77
IMR        0.24    0.27       0.16      0.92 1.00    0.80
LitRate    0.20    0.22       0.13      0.77 0.80    1.00
```

Figure 7.12 : Restricted or the estimated correlation matrix of sem model-2

Non – restricted or the original correlation matrix

```
            GDP Imports Population LE_Male  IMR LitRate
GDP        1.00    0.94       0.58    0.15 0.15    0.13
Imports    0.94    1.00       0.56    0.28 0.24    0.19
Population 0.58    0.56       1.00    0.00 0.00   -0.02
LE_Male    0.15    0.28       0.00    1.00 0.87    0.72
IMR        0.15    0.24       0.00    0.87 1.00    0.80
LitRate    0.13    0.19      -0.02    0.72 0.80    1.00
```

Figure 7.13: Non-Restricted or the actual correlation matrix of sem model-2

7.4.5. Root Mean square error - Model 2 :

The root mean square error (RMSE) measures the average difference between the predicted correlation matrix and the actual correlation matrix.

The root mean square error value for our estimated and original matrix is 0.075 which is less than 0.1 , so our model is acceptable.

Now we estimate other indexes that are used to test the goodness fit of the model. We use GFI (Goodness- of fit index , AGFI (Adjusted Goodness of fit index) , SRMR (Square root Mean Residual) – which is the root mean square error computed above.

```
Model Chisquare =  28   Df =  8 Pr(>Chisq) = 0.00054
Goodness-of-fit index =  0.95
Adjusted goodness-of-fit index =  0.87
SRMR =  0.069

Normalized Residuals
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -2.28   -0.99   -0.02   -0.52    0.00    0.41

R-square for Endogenous Variables
LE_Female        IMR    LitRate        GDP    Exports Population
     0.89       0.96       0.66       0.81       1.02       0.37

Parameter Estimates

Iterations =  31
```

Figure 7.14: Parameter Estimations for goodness of fit- Model2

From the above table , we can say that the GFI is 0.94 , AGFI is 0.86 and SRMR is 0.051. Any model with GFI greater than 0.9 , AGFI greater than 0.8 and SRMR less than 0.1 can be considered a good model. So, we can say that our model is good and acceptable.

To visually represent our model, we use path diagram which explains the correlation between manifest variables, the correlation between the factors and the variances between the manifest variables (uniqueness) and the covariance between the factors.

7.4.6. Path Diagram for our specified CFA model 2

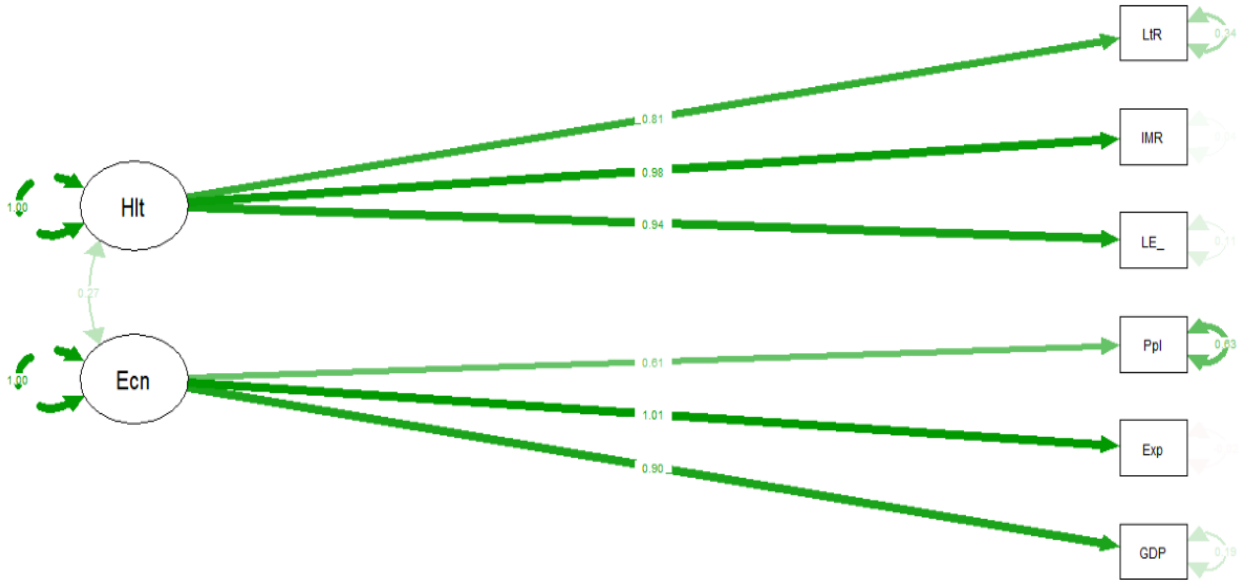

Figure 7.15 : Path Diagram for CFA model

From the results of the Confirmatory Factor Analysis , based on the assumption that the latent factors health (living or social) conditions of the countries depends on the manifest variables like the life Expectancy of the male (female) , the infant mortality rate and the literacy rate. The Economic indicators may vary depending on the manifest variables like the Gross Domestic Product ,the Imports (Exports) and the population, we can conclude that the assumption holds true, and the social and economic factors of a country depends on the mentioned manifest variables.

8. Conclusion

In conclusion, this multivariate analysis of country-level economic, environmental, and social indicators provided valuable insights into the relationships between different factors of development across nations. Through extensive data cleaning and preprocessing, we ensured a clean and consistent dataset suitable for multivariate techniques.

Dimension reduction via principal component analysis revealed the key underlying dimensions in the data, reducing its complexity while retaining over 75% of the information. The top principal components highlighted differences related to socioeconomic development levels, economic productivity, environmental sustainability, and cultural diversity. Visualizing the relationships through biplots and 3D scatterplots aided effective interpretation.

Both clustering algorithms and dimension reduction techniques identified meaningful groupings of countries based on similar profiles across indicators. K-Means clustering outperformed other methods in distinguishing three robust clusters representing higher socioeconomic status nations, lower status nations, and those with stagnating economies.

Confirmatory factor analysis of the hypothesized relationships between latent and manifest variables confirmed that a nation's health and living standards are influenced by life expectancy, infant mortality, and literacy rates. Similarly, economic indicators depend on GDP, international trade levels and population size. Model fit statistics validated the acceptability of this theoretical CFA model.

Overall, this comprehensive analysis of the multifaceted drivers of national progress through multiple multivariate lenses provided a well-rounded understanding rarely achievable through single variable or bivariate analyses. Policymakers seeking to balance economic growth with societal welfare can utilize these types of insights to inform decision making and target development efforts more precisely. The approach followed in this study can be extended to other domains as well to uncover meaningful patterns from complex, real-world multivariate data.

9. Appendix

Description of the Variables:

ID	Variable Name	Variable Description
1	Country	Country Names
2	Population	Population of countries , year 2022[2]
3	LE_Male	Life Expectancy of Male
4	LE_Female	Life Expectancy of Female
5	UnEmpRate	Un employment rate
6	SexRatio	Sex ratio (how many males for 100 female)
7	IMR	Infant Mortality Rate
8	LitRate	Literacy Rate[4]
9	CDI	Cultural Diversity Index
10	GDP	Gross Domestic Product[2]
11	PerCap	Per Capita Income[2]
12	InfRate	Inflation Rate[3]
13	Exports	Exports (Million US$)
14	Imports	Imports (Million US$)
15	Ecn_Agri	Economy percent from Agricultural sector (% of GDP)
16	Ecn_Ind	Economy percent from Industrial sector (% of GDP)
17	Ecn_Serv	Economy percent from Service sector (% of GDP)
18	NDRI	National Disaster Recovery Index
12	NDRI_Cat	Categorized National Disaster Recovery Index
13	CO2	Carbon Dioxide Emissions Value
14	AQI	Air Quality Index
15	AQI_Cat	Categorized Air Quality Index
16	WQI	Water Quality Index
17	WQI_Cat	Categorized Water Quality Index

10. References:

[1] Suddala, Rajkumar. (2018). Country Statistics- UNData. Kaggle. Retrieved from https://www.kaggle.com/datasets/sudalairajkumar/undata-country-profiles

[2] Worldometer.(2023).GDP.GDP by Country. Retrieved from https://www.worldometers.info/gdp/gdp-by-country/

[3] Trading Economics. (2023). Inflation Rate by Country. Retrieved from https://tradingeconomics.com/country-list/inflation-rate-

[4] Wise Voter. (2023). Country Rankings. Literacy Rate by Country. Retrieved From https://wisevoter.com/country-rankings/literacy-rate-by-country/

[5] Worldometer.(2023).Demographics. Life Expectancy of the world population. Retrieved from https://www.worldometers.info/demographics/life-expectancy/

[6] Silhouette Coefficient. OpenAI. (Year). GPT-3.5 architecture. Retrieved from https://www.openai.com/

[7] Scikit-learn developers. (2023). sklearn.metrics.silhouette_score. Retrieved fromhttps://scikitlearn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html#:~:text=The%20Silhouette%20Coefficient%20is%20calculated,is%20not%20a%20part%20of