|  |
| --- |
| Photo displaying partial image of two pie charts on a canvas-textured page |
| Sustainable Development and Social well being  A Multivariate Analysis of Economic, Environment and Social domains |
| |  |  | | --- | --- | |  |  | |  |  | |

**Table of Contents**

Introduction -------------------------------------------------------------------------------------------------- 1

1.Data Set ---------------------------------------------------------------------------------------------------- 2

2. Multivariate Analysis ------------------------------------------------------------------------------------ 2

3. Data Cleaning Techniques ------------------------------------------------------------------------- (3 -5)

3.1. Adding of new Columns to the dataset ----------------------------------------------------- 3

3.2. Handling Missing Values --------------------------------------------------------------------- 4

3.3. Conversion of non-numeric data to numeric data ----------------------------------------- 4

3.4. Direction Changing --------------------------------------------------------------------------- 4

3.5. Detecting Outliers ----------------------------------------------------------------------------- 4

3.6. Scaling and Preparing the final data -------------------------------------------------------- 5

4. Data Visualizations ---------------------------------------------------------------------------------- (6 -9)

4.1. Mahalanobis distances ------------------------------------------------------------------------ 6

4.2. Box Plot – Univariate ------------------------------------------------------------------------- 6

4.3. Bivariate Boxplot ------------------------------------------------------------------------------ 7

4.4. Bubble Plot ------------------------------------------------------------------------------------- 7

4.5. Scatterplot Matrix ----------------------------------------------------------------------------- 8

4.6. Scatterplot 3D --------------------------------------------------------------------------------- 9

4.6. GG plot ----------------------------------------------------------------------------------------- 9

5. Dimension Reduction Techniques -------------------------------------------------------------- (10 -15)

5.1. Principal Component Analysis--------------------------------------------------------------- 10

5.2. Results of the PCA ---------------------------------------------------------------------------- 10

5.3. PCA Biplot ------------------------------------------------------------------------------------- 11

5.4. PCA- 3D Plot---------------------------------------------------------------------------------- 12

5.5. Visually representing PCA on a map ------------------------------------------------------- 14

5.6. Multi – Dimensional Scaling ---------------------------------------------------------------- 15

6. Cluster Analysis ----------------------------------------------------------------------------------- (18 - 30)

6.1. Hierarchical Clustering ---------------------------------------------------------------------- 18

6.2. K – Means Clustering ------------------------------------------------------------------------ 22

6.3. Model – Based Clustering -------------------------------------------------------------------- 25

6.4 Evaluating the clustering outcomes -------------------------------------------------------- 29

7. Confirmatory Factor Analysis ------------------------------------------------------------------- (31 – 40)

7.1. Factor Analysis ------------------------------------------------------------------------------- 31

7.2. UN Statistics Data Set ----------------------------------------------------------------------- 31

7.3. Confirmatory Factor Analysis -------------------------------------------------------------- 31

7.3.1. Requirements for CFA ------------------------------------------------------------ 31

7.3.2. Results for EFA assuming 3 factors -------------------------------------------- 32

7.4. CFA model ------------------------------------------------------------------------------------- 35

7.4.1. Model1 ------------------------------------------------------------------------------ 35

7.4.2. Root Mean square error – Model1---------------------------------------------- 36

7.4.3. Path Diagram of Model1 --------------------------------------------------------- 37

7.4.4. Model2 ------------------------------------------------------------------------------ 38

7.4.5. Root Mean square error – Model2 ---------------------------------------------- 39

7.4.6. Path Diagram of Model2 --------------------------------------------------------- 40

8. Conclusion ------------------------------------------------------------------------------------------------ 41

9. Appendix -------------------------------------------------------------------------------------------------- 42

10. References -----------------------------------------------------------------------------------------------43

**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 the individually 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.



**ECONOMIC**

* GDP
* Per Capita Income
* Inflation Rate
* Exports of Goods
* Imports of Goods
* Sector Wise GDP Division

- Agriculture

- Industry

- Service

**ENVIRONMENTAL**

* Natural Disaster Recovery Index
* NDRI Category
* Co2 Emissions
* Air Quality Index
* AQI Category
* Water Quality Index
* WQI Category

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 are 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. Conversion of Non-Numerical data to numerical data:**

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.

A screenshot of a computer screen

Description automatically generated

Figure 3.1 : Top 6 observations of the scaled dataset.

A screenshot of a computer code

Description automatically generated

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**

A green line graph with numbers

Description automatically generated

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.

A yellow rectangular object with black lines

Description automatically generated

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.

A diagram of a diagram

Description automatically generated with medium confidence

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.

A diagram of circles

Description automatically generated

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.

A group of black and white squares

Description automatically generated

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

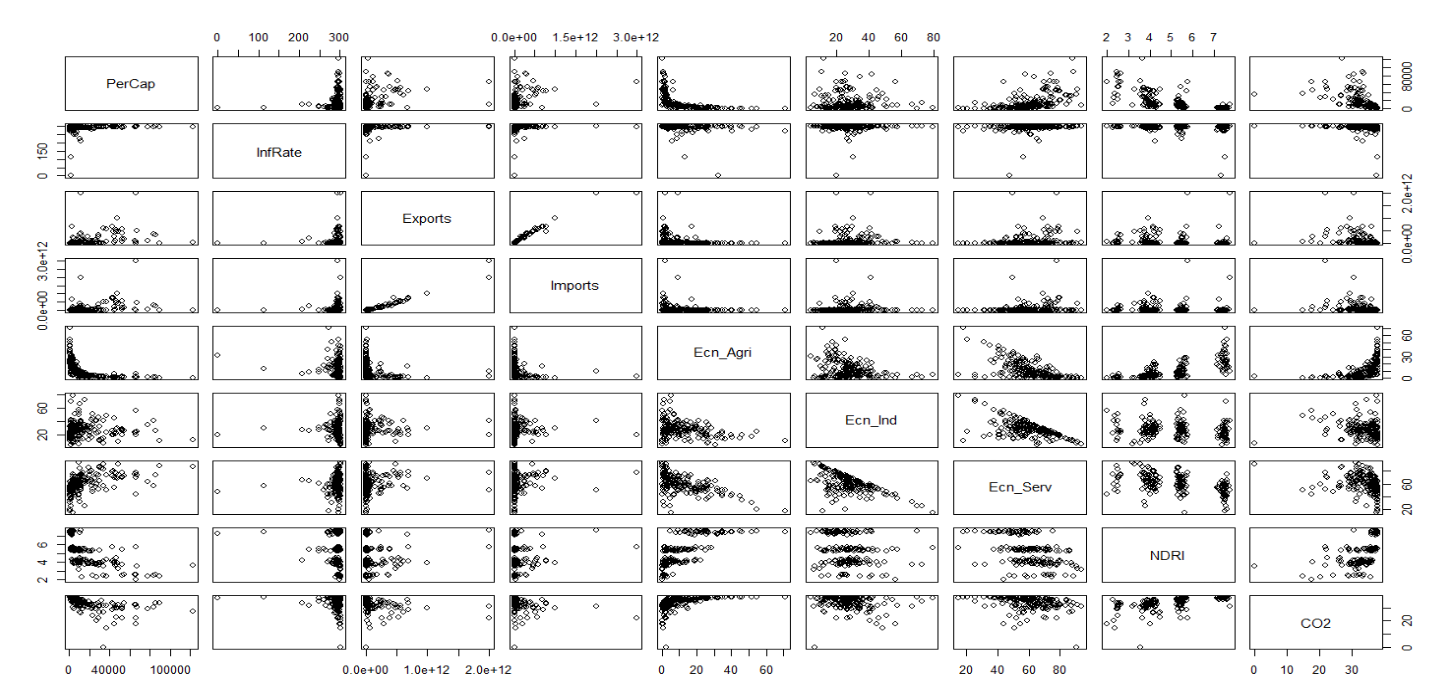


Figure 4.6. Scatterplot Matrix for the last 9 variables in the dataset

**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.

A graph with a line graph

Description automatically generated with medium confidence

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.

**A map of the world

Description automatically generated**

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:**

The PCA identified 6 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.

A close-up of several graphs

Description automatically generated

Figure 5.1: 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.

A graph of a graph with colored dots

Description automatically generated with medium confidence  
Figure 5.2 : 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.

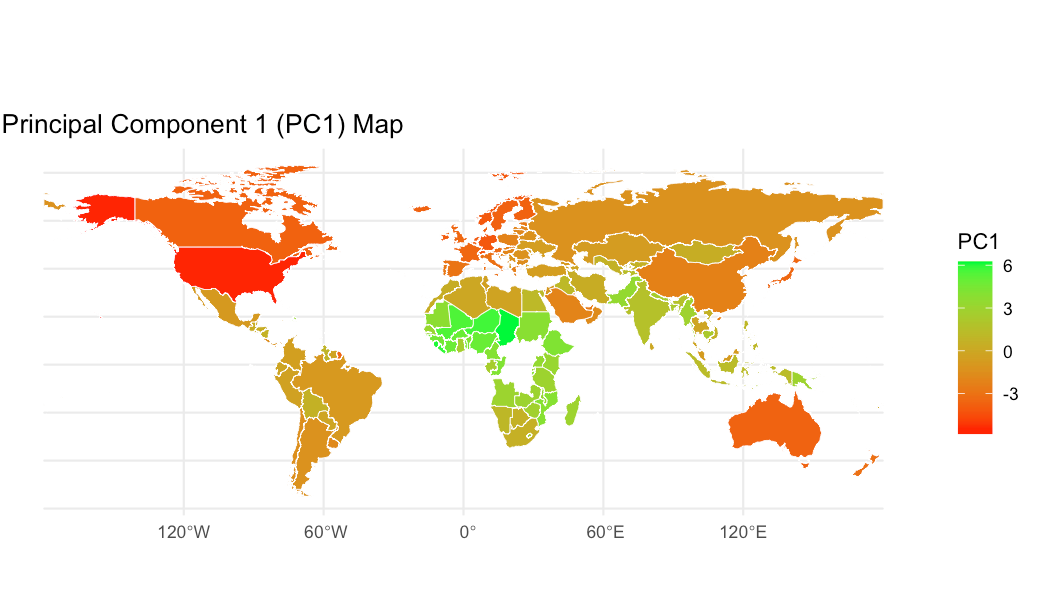
  PC2 deals with economic productivity/income as discussed earlier. Countries towards the right side could be more developed/industrialized.

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**



A map of the world

Description automatically generated  
A map of the world with different colored countries/regions

Description automatically generated

Figure 5.3: 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.

A table with numbers and symbols

Description automatically generated

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

A graph with red dots

Description automatically generated

Figure 5.4: 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.

A graph of a person with a hand holding a hand

Description automatically generated with medium confidenceFigure 5.5 : 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.

A close-up of a number

Description automatically generated

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.

A graph with text and numbers

Description automatically generated with medium confidenceFigure 5.6 : 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.

* 1. **Clustering Techniques:**
     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.

* + 1. **Hierarchical Clustering - Single Linkage**

Single Linkage – Choosing the smallest distance between the groups.

A graph of a person

Description automatically generated  
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.

* + 1. **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

A graph of a graph

Description automatically generated

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.

A diagram of a cluster of data

Description automatically generated

Figure 6.7: Dendrogram - Hierarchical Complete linkage Clustering

* + 1. **Hierarchical Clustering - Average Linkage**

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

A graph of a graph

Description automatically generated

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.

* 1. **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.

A graph of a number of clusters

Description automatically generated

Figure 6.11 : Scree plot for K- Means clustering.

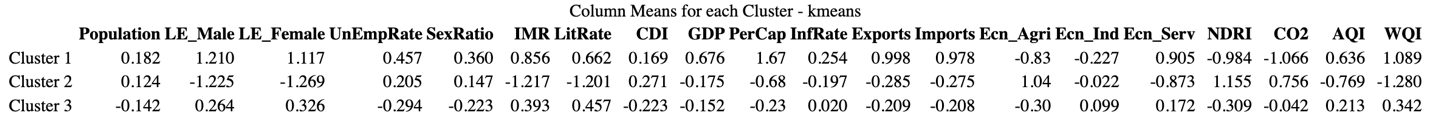
Based on the scree plot, set k = 3 as the cut point and we get the below table as clusters outcome.

|  |  |  |
| --- | --- | --- |
| **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.

  
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".
  + 1. **Interpreting K-Means based on PCA**.

A graph of different colored lines

Description automatically generated

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.

A map of the world

Description automatically generated

Figure 6.15 : Map showing the K-Means clustering.

* 1. **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.

A graph of different colored lines

Description automatically generated

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.

A number of numbers and a few words

Description automatically generated with medium confidence

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 .

A screenshot of a computer

Description automatically generated  
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.

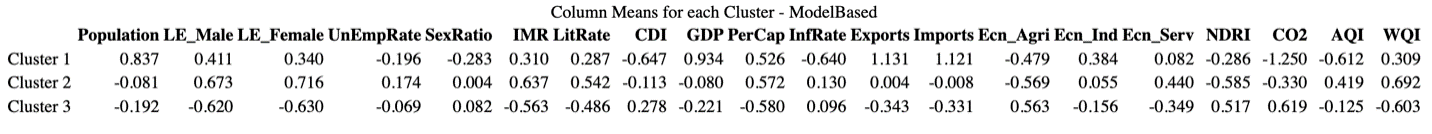


Table 6.20: ColumnMeans, Model based Clustering.

* 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.**

A graph showing a number of different countries/regions

Description automatically generated with medium confidence

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.

A map of the world

Description automatically generated

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.
  + 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.

A white background with black text

Description automatically generated

A graph with numbers and lines

Description automatically generated

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

* + 1. **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.

A screenshot of a computer

Description automatically generated

A graph with numbers and lines

Description automatically generated

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

1. **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.**

* 1. **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

* 1. **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.

* + 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.

* + 1. **Results for the Exploratory factor analysis assuming 3 factors.**

A screenshot of a computer

Description automatically generatedFigure 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.

A screen shot of a number

Description automatically generated

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.

A white background with black text

Description automatically generated

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.

A screenshot of a computer

Description automatically generated

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

* 1. **CFA Model :**
     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.

A screenshot of a graph

Description automatically generated

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 :**

Confidence Interval = Estimate ± CriticialValue \* 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

A screenshot of a computer screen

Description automatically generated

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

# Non – restricted or the original correlation matrix

A screenshot of a computer screen

Description automatically generated

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

* + 1. **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.

A white background with black text

Description automatically generated

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.

* + 1. **Path Diagram for our specified CFA model 1**

A green lines in a line

Description automatically generated with medium confidence

Figure 7.10: Path Diagram for CFA model – Model1

* + 1. **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.

A screenshot of a computer

Description automatically generated

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 :**

Confidence Interval = Estimate ± CriticialValue \* 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

A screenshot of a computer screen

Description automatically generated

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

# Non – restricted or the original correlation matrix

A screenshot of a computer screen

Description automatically generated

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

* + 1. **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.

A white background with black text

Description automatically generated

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.

* + 1. **Path Diagram for our specified CFA model 2**

A green lines on a white background

Description automatically generated

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.

1. **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.

1. **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](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 from <https://scikitlearn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html#:~:text=The%20Silhouette%20Coefficient%20is%20calculated,is%20not%20a%20part%20of>.