Yash Gupta

Multivariate Analysis of World Data

1 Introduction

In our ever-connected world, Humanity has spread all over the earth, creating different societies with distinct traditions, rituals, and features. However, there is an underlying thread of shared characteristics that unite us as a species despite our many differences. With more than 190 countries recognized by the UN, each with a unique character, this research aims to explore the nuances that make these countries distinct yet interdependent on each other.

To promote a more inclusive and peaceful global community, it is essential that we recognize the subtleties of our differences and find the things that bring us together.

We aim to uncover the underlying relationships between different socio-economic and environmental variables through a comprehensive examination of multivariate data. Identifying and making use of correlations between variables, from economic and health indicators to environmental and social indices, it is possible to find trends, highlight dependencies, and extract knowledge that can guide international projects, policy decisions, and decision-making.

2 Data Collection

Data used for this analysis was extracted from the **Gapminder Foundation** [1]website. Gapminder Foundation is a non-profit Swedish-based organisation which uses statistics to promote sustainable global development and contains comprehensive information about many features ranging from economic to social features of various countries. Of all available features few relevant features were picked from Health, Economic and Environmental indicators for this analysis.

Health indicators include:

(a) Newborn mortality rate per 1000
(b) Babies per woman (total fertility)
(c) DTP3 immunized (% of one-year-olds)
(d) Government health spending per person (US$)

Environmental indicators include:

(a) Cell phones (per 100 people)
(b) Electricity use, per person
(c) CO2 emissions yearly (tonnes per person)

Economic indicators include:

(a) Gross National Income (GNI) per capita, PPP, current international
(b) Children and elderly (per 100 adults)
(c) Corruption Perception Index (CPI)
(d) GDP/capita growth over the next 10 years
(e) Exports (% of GDP)
(f) Foreign investment inflows (direct, net % of GDP)
(g) Females aged 15+ labour force participation rate (%)

Every one of these indicators offers insightful information on the general health, progress, and sustainability of a country. The set of data aims to represent the complexity of health outcomes, environmental sustainability, economic performance, and country development.

Each of them consists of entire datasets of years and years of data, but for our analysis, we have picked the most latest and relevant data from each file and combined them in a single CSV document as shown in figure 1.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1			HEALTH							ECONOMY				ENVIRONMENT		
2	COUNTRY		Newborn mortality rate per 1000 (2015)	Babies per woman (total fertility)(2024)	DTP3 immunized (% of one-year-olds)(2019)	Government health spending per person (US$)(2010)	Gross National Income (GNI) per capita (2022)	Children and elderly (per 100 adults)(2024)	Corruption Perception Index (CPI)(2017)	GDP/capita growth over the next 10 years(2001)	Exports (% of GDP)(2022)	Foreign investment inflows (direct, net % of GDP)(2021)	Females aged 15+ labour force participation rate (%)(2019)	Cell phones (per 100 people)(2021)	Electricity use, per person(2014)	CO2 emissions yearly (tonnes per person)(2018)
3	Afghanistan		35.5	3.53	66	4.39		81.8	15	7.8		0.144	21.8	56.6		0.254
4	Albania		6.2	1.7	99	94	18.3k	50.5	38	4.44	37.4	6.8	52.5	92.3	2310	1.59
5	Algeria		15.5	2.39	91	139	12.9k	58.7	33	2.11	35.3	0.532	17	106	1370	3.69
6	American Samoa							53.9			47					
7	Andorra							39.3		-0.511				119		6.12
8	Angola		1.4	5.14	99	2170	6450	89.5	19	7.84	43.4	-6.55	76.1	44.4	310	1.12
9	Anguilla							39.1		1.72						
10	Antigua and Barbuda		48.7	1.97	57	102	25.6k	41.7		0.563	54.7	15.3		197		5.88
11	Argentina		4.9	2.18	95	490	26k	53	39	4.55	16.3	1.41	51.3	130	3070	4.41
12	Armenia		6.3	1.58	86	405	18.1k	52.7	35	7.43	50	2.64	42.9	129	1980	1.89
13	Aruba		7.4	1.79	92	54.2	47.1k	48.9		-1.63	83.1	4.3		132		
14	Australia		2.2	1.8	95	3250	60.8k	54.7	77	1.61	25.4	1.59	60.8	105	10.1k	16.9
15	Austria		2.1	1.57	85	3840	67.8k	53.8	75	1.3	62.1	2.57	55.2	122	8360	7.75
16	Azerbaijan		18.2	1.97	94	67.3	16.6k	43.8	31	12.8	60.1	-3.11	63.3	105	2200	3.7
17	Bahamas		6.9	1.72	86	791	38.8k	38.2	65	-0.671	37.5	3.32	70.1	97.4		4.7
18	Bahrain		1.1	1.86	99	634	58.5k	31.3	36	0.605		4.53	45.3	131	20k	19.8
19	Bangladesh		23.3	1.9	98	58	7690	46.3	28	4.4	12.9	0.414	36.4	109	317	0.531
20	Barbados		8	1.81	90	652	17.5k	50.9	68	1.14	35.1	4.85	61.8	113		4.48
21	Belarus		1.9	1.76	98	248	21.7k	52.9	44	7.85	63.7	1.77	57.6	123	3690	6.93
22	Belgium		2.2	1.82	98	3450	66.5k	57.8	75	0.905	95.7	4	49	101	7710	8.69
23	Belize		8.3	2.27	98	150	10.7k	47.6		0.893	52.4	5.18	50	66		1.5
24	Benin		31.8	4.45	76	15	4020	82.4	39	0.835	21.7	1.96	68.8	98	97.1	0.623
25	Bermuda						98.6k	56.5		1.71	52.3	0.697		106		
26	Bhutan		18.3	1.83	97	94.2		37.9	67	6.33		0.245	59.5	100		1.61
27	Bolivia		19.6	2.6	75	60.6	9460	54.1	33	2.23	32.6	1.44	63.7	99.6	728	1.96

Figure 1: Merged dataset.

The collection and merging of data were the most time-consuming part of this analysis. However, once organized, the cumulated dataset contained health, economic and environmental information in 14 columns of 236 countries. Compiled Dataset was then uploaded to Kaggle for public use.

Still, lot of countries had incomplete information and a lot of cleaning and data preparation was necessary before any kind of analysis was possible.

3 Data Cleaning and Preparation

Out of 236 countries, several countries had a lot of missing values, this might be because of small population size, Political Instability, Limited infrastructure, isolation from other nations, Resource constraints or Data privacy issues and in order to create a generalised analysis for the current state of collective global nature, Deletion of such entries was necessary.

So, any country with more than 5 missing values was removed from the analysis i.e. 46 nations were removed.

```
df_clean <- df[complete.cases(df) | rowSums(is.na(df)) <= 5, ]
```

Next, two columns **"Electricity use, per person"** and **"Gross National Income (GNI) per capita, PPP, current international"** which were supposed to be in numeric datatype contained character "k" signifying "times 1000", so a function to convert these columns into numeric datatype was created as shown in figure 2.

	`(GNI) per capita` <chr>	`(GNI) per capita` <dbl>	Electricity <chr>	Electricity <dbl>
1	NA	NA	NA	NA
2	18.3k	18300	2310	2310000
3	12.9k	12900	1370	1370000
4	6450	6450000	310	310000
5	25.6k	25600	NA	NA
6	26k	26000	3070	3070000
7	18.1k	18100	1980	1980000
8	47.1k	47100	NA	NA
9	60.8k	60800	10.1k	10100
10	67.8k	67800	8360	8360000

Figure 2: Conversion to Numeric datatype

In column **"GDP/capita growth over the next 10 years**" and **"Foreign investment inflows (direct, net % of GDP)"**, negative values were denoted by a negative symbol not supported by R, so a function to replace all of the negative symbol to "-" symbol recognized by R was created and implemented i.e.

$$-0.567 => -0.567$$

Now, The Correlation and relation between every variable is displayed in figure 3.

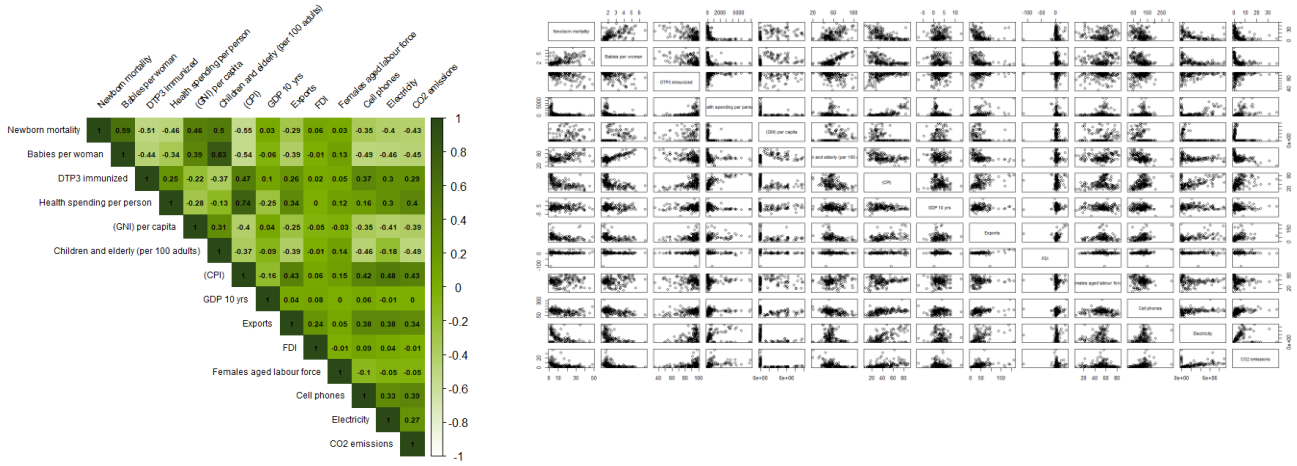


Figure 3: (a)Correlation Plot (b)Pair Plot of all features

The most positively correlated variables are: Children and elderly (per 100 adults) and Babies per woman with a correlation value: 0.8311021 and the Most negatively correlated variables are: Corruption Perception Index (CPI) and Newborn mortality rate with a correlation value: -0.5526628.

Summary statistics of dataset is presented in figure 4.

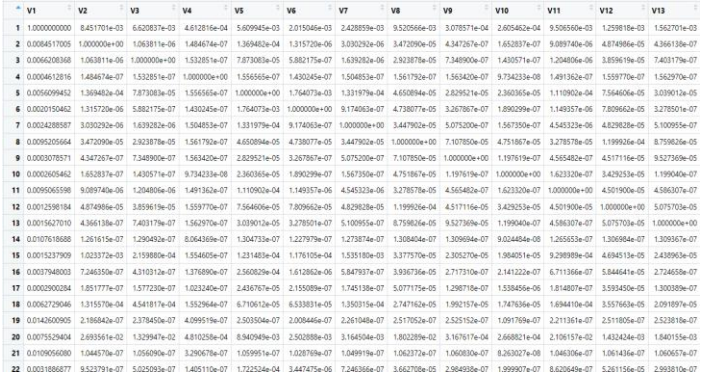
```
> summary(df_clean)
 Newborn mortality Babies per woman DTP3 immunized Health spending per person (GNI) per capita  Children and elderly (per 100 adults)     (CPI)
 Min.   : 0.700    Min.   :1.260    Min.   :35.0   Min.   :   2.09            Min.   :  10500   Min.   : 20.50                       Min.   : 9.00
 1st Qu.: 4.375    1st Qu.:1.745    1st Qu.:85.0   1st Qu.:  33.80            1st Qu.:  21800   1st Qu.: 49.70                       1st Qu.:29.00
 Median :10.200    Median :2.120    Median :93.0   Median : 160.00            Median :  53300   Median : 54.80                       Median :38.00
 Mean   :13.651    Mean   :2.568    Mean   :88.2   Mean   : 690.91            Mean   :1648992   Mean   : 58.58                       Mean   :42.98
 3rd Qu.:21.825    3rd Qu.:3.335    3rd Qu.:97.0   3rd Qu.: 574.00            3rd Qu.:2730000   3rd Qu.: 66.95                       3rd Qu.:56.00
 Max.   :48.700    Max.   :6.710    Max.   :99.0   Max.   :6910.00            Max.   :9830000   Max.   :104.00                       Max.   :89.00
 NA's   :2         NA's   :3        NA's   :6      NA's   :9                  NA's   :13                                             NA's   :12
   GDP 10 yrs        Exports            FDI          Females aged labour force  Cell phones      Electricity      CO2 emissions
 Min.   :-7.140   Min.   :  1.57   Min.   :-113.000   Min.   : 6.04             Min.   : 19.40   Min.   :  10100   Min.   : 0.0243
 1st Qu.: 0.872   1st Qu.: 26.10   1st Qu.:   1.137   1st Qu.:44.83             1st Qu.: 84.85   1st Qu.: 313500   1st Qu.: 0.6610
 Median : 2.460   Median : 41.00   Median :   2.810   Median :54.05             Median :111.50   Median :1680000   Median : 2.5300
 Mean   : 2.602   Mean   : 47.87   Mean   :   3.622   Mean   :52.07             Mean   :108.87   Mean   :2502949   Mean   : 4.4279
 3rd Qu.: 4.250   3rd Qu.: 57.95   3rd Qu.:   4.973   3rd Qu.:61.02             3rd Qu.:131.75   3rd Qu.:3940000   3rd Qu.: 5.8600
 Max.   :12.800   Max.   :211.00   Max.   :  37.200   Max.   :83.90             Max.   :319.00   Max.   :9050000   Max.   :38.0000
 NA's   :1        NA's   :43       NA's   :10         NA's   :10                                  NA's   :51        NA's   :1
```

Figure 4: Summary statistics of the dataset

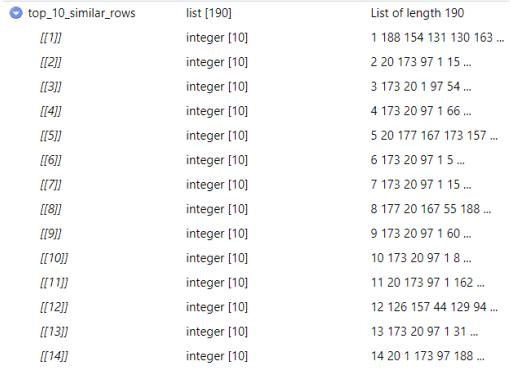
Still with so many missing values, a proper analysis cannot be done till they are present in the dataset. There are three choices for dealing with them [2]:

1) **Removing missing values:**
   Simple and straightforward approach but this will lead to the loss of a lot of data.
2) **Data imputation:**
   Retains more data than removing missing values and allows for the utilization of existing data but imputation methods can introduce bias, every country is unique and using stats of entire data does not justify this approach. The unique circumstances of countries like war-torn Afghanistan shouldn't be clubbed with a developed nation like the United States.
3) **Finding similar countries and imputing mean:**
   This approach takes into account the similarities between countries and can potentially provide more accurate imputations compared to using the overall mean of the entire dataset.

For this analysis, the third approach is most appropriate because it preserves the uniqueness of each country. So, a similarity matrix was created using Euclidean distance for each country and the top 10 most similar countries were found. Using the complete values from these selected countries, missing values of respective countries were found and cleaned data now looks like figure 5.



(a)



(b)

```
> summary(df_clean)
 Newborn mortality Babies per woman DTP3 immunized  Health spending per person (GNI) per capita  Children and elderly (per 100 adults)     (CPI)
 Min.   : 0.70     Min.   :1.260    Min.   :35.00   Min.   :   2.09            Min.   :  10500   Min.   : 20.50                       Min.   : 9.00
 1st Qu.: 4.40     1st Qu.:1.760    1st Qu.:84.25   1st Qu.:  35.98            1st Qu.:  22950   1st Qu.: 49.70                       1st Qu.:29.00
 Median :10.70     Median :2.150    Median :92.00   Median : 152.96            Median :  60400   Median : 54.80                       Median :37.00
 Mean   :13.68     Mean   :2.563    Mean   :88.03   Mean   : 664.27            Mean   :1736128   Mean   : 58.58                       Mean   :41.95
 3rd Qu.:21.77     3rd Qu.:3.325    3rd Qu.:97.00   3rd Qu.: 543.75            3rd Qu.:3136417   3rd Qu.: 66.95                       3rd Qu.:55.00
 Max.   :48.70     Max.   :6.710    Max.   :99.00   Max.   :6910.00            Max.   :9830000   Max.   :104.00                       Max.   :89.00
   GDP 10 yrs         Exports            FDI          Females aged labour force  Cell phones      Electricity      CO2 emissions
 Min.   :-7.1400   Min.   :  1.57   Min.   :-113.000   Min.   : 6.04             Min.   : 19.40   Min.   :  10100   Min.   : 0.0243
 1st Qu.: 0.8772   1st Qu.: 28.19   1st Qu.:   1.215   1st Qu.:44.92             1st Qu.: 84.85   1st Qu.: 622081   1st Qu.: 0.6637
 Median : 2.5000   Median : 40.90   Median :   2.815   Median :53.35             Median :111.50   Median :1454750   Median : 2.4950
 Mean   : 2.6033   Mean   : 45.83   Mean   :   3.624   Mean   :51.78             Mean   :108.87   Mean   :2147436   Mean   : 4.4166
 3rd Qu.: 4.2325   3rd Qu.: 52.70   3rd Qu.:   4.957   3rd Qu.:60.80             3rd Qu.:131.75   3rd Qu.:3035000   3rd Qu.: 5.8325
 Max.   :12.8000   Max.   :211.00   Max.   :  37.200   Max.   :83.90             Max.   :319.00   Max.   :9050000   Max.   :38.0000
```

(c)

Figure 5: (a) Similarity matrix of each country (b) Top 10 most similar countries to one another (c)Summary statistics of cleaned data

4 Modelling

Principal Component Analysis:

A dimensionality reduction approach called principal component analysis (PCA) [3] was used to find patterns in data.

- It was used to reduce the dimensionality of a dataset made up of variables that are connected with one another while maintaining the variance found in the data set by converting

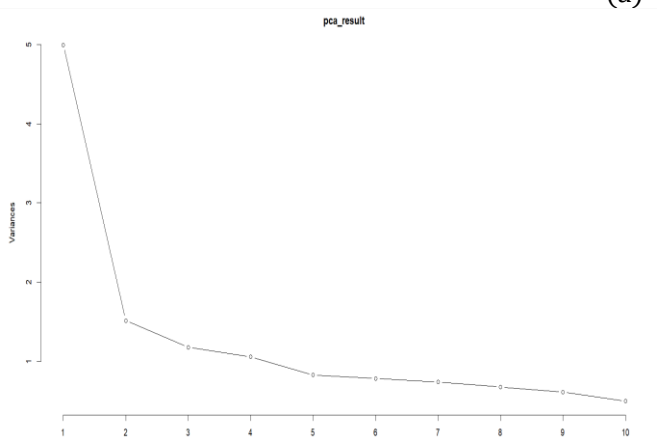
the original variables into a new collection of uncorrelated variables known as **principal components**.

- The principal components are uncorrelated with each other. Thus, each individual aspect of the variation in the data was captured by each principal component.
- Maximum variation in the data is explained by the first main component. Orthogonal to the preceding components, each succeeding principal component represents the largest residual variance.
- By modelling the data with fewer primary components and reducing information loss, PCA was utilized for data compression.

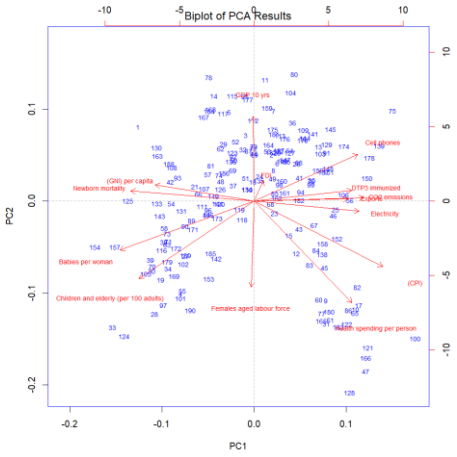
With a standard deviation of 2.2346, PC1 explains the dataset's variation the most.
The percentage of variance explains the percentage of overall variance that each primary component accounts for. PC1 accounts for 35.67% of the variance overall. The entire variance explained by the major components is represented by the cumulative proportion, the sum of PC1 through PC8 accounts for 84% of the variation.

```
> summary(pca_result)
Importance of components:
                          PC1    PC2     PC3     PC4    PC5     PC6     PC7    PC8     PC9    PC10    PC11    PC12    PC13    PC14
Standard deviation     2.2346 1.2310 1.08577 1.02898 0.9104 0.88502 0.86001 0.8231 0.78321 0.70702 0.6776 0.59436 0.43557 0.33009
Proportion of Variance 0.3567 0.1082 0.08421 0.07563 0.0592 0.05595 0.05283 0.0484 0.04382 0.03571 0.0328 0.02523 0.01355 0.00778
Cumulative Proportion  0.3567 0.4649 0.54911 0.62474 0.6839 0.73989 0.79272 0.8411 0.88493 0.92064 0.9534 0.97867 0.99222 1.00000
```

(a)



(b)



(c)

Figure 6: (a) summary of principal components (b)Scree Plot (c)Biplot

In order to interpret the results of PCA, a scree plot and Biplot were generated.

- The number of primary components to be retained in the study is decided using **the scree plot**. It presents each primary component's eigenvalues in descending order.
- The relationship between the original variables and the principal components is shown in a single plot by a **biplot**. The space bounded by the principal components of a biplot includes the original variables as well as the observations, or data points that makes it easier for you to see how different variables affect the principal components and how these components relate to one another in the data.

Hierarchical Clustering:

Hierarchical clustering [4]is a method used to group similar objects into clusters based on their pairwise similarities or dissimilarities.

- Agglomerative clustering was used, in which each data point forms its own cluster at first and is gradually combined with the clusters that are the most similar to each other until only one cluster is left.
- Complete linkage was used in the procedure, where the maximum distance between any pair of points in two clusters is used to define the distance between them. This method guarantees that the largest dissimilarity between the members of a cluster will be used to merge it.
- Ultimately, the hierarchical structure of the clusters was depicted using a dendrogram. The dendrogram shows the way clusters merge at each algorithmic step as shown in figure 7.

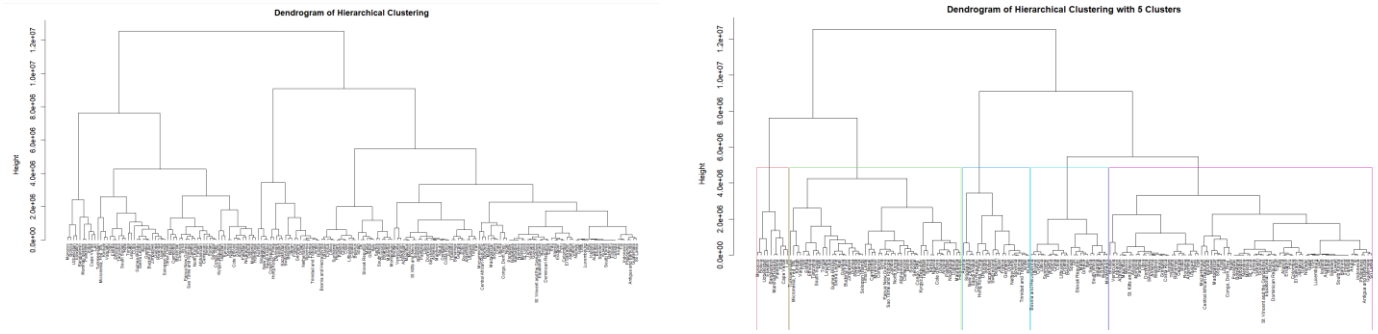
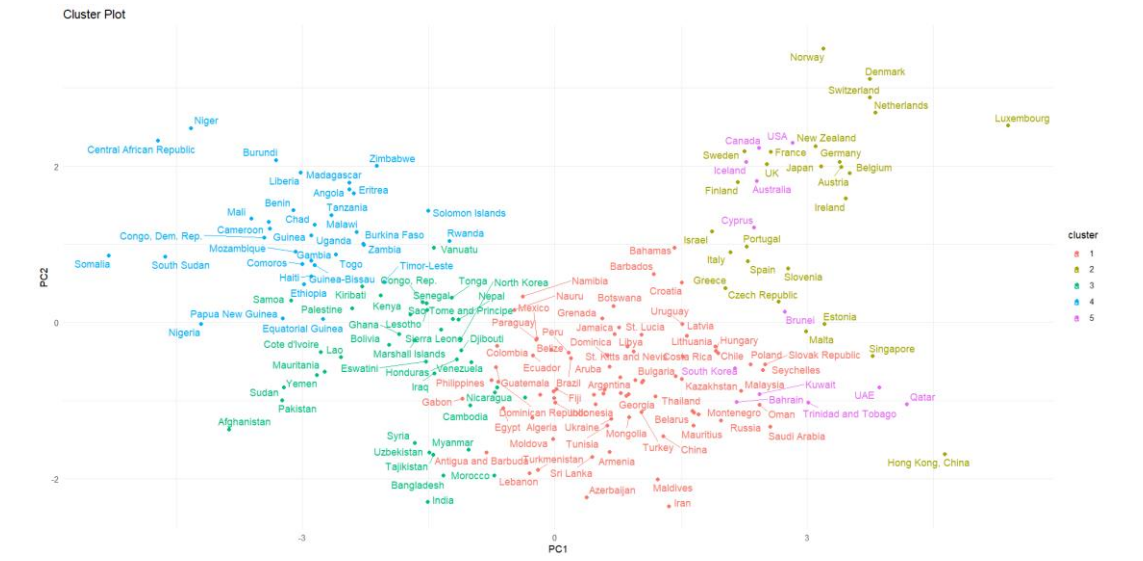


Figure 7: Dendrogram of Hierarchical Clustering

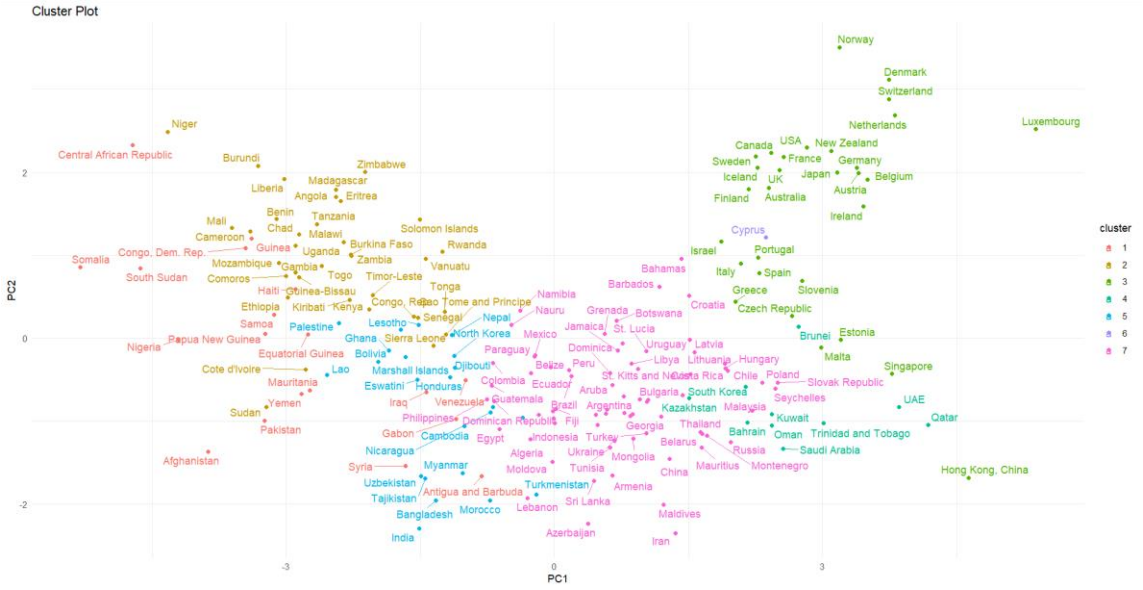
K-means clustering:

K-means clustering [5] is an unsupervised learning method that divides data into k different groups according to how similar their features are.

- The algorithm changes the centroids based on the mean of the data points assigned to each cluster after iteratively assigning each data point to the closest cluster centroid.
- When convergence is achieved, the algorithm generates clusters with centroid positions.
- The k-means method was applied to the reduced-dimensional space defined by the first two principal components.
- This method makes it easier to recognize different clusters and the spatial interactions between them, which helps to explain the underlying patterns in the dataset.



(a)



(b)

Figure 8: Clustering of countries using K means clustering technique (a) 5 clusters (b) 7 clusters

5 Conclusion

To sum up, this investigation highlights the complex interactions between socioeconomic and environmental factors in over 190 countries, providing insight into the diverse characteristics of the world community. We started by carefully gathering and preparing the data by hand to guarantee the accuracy and consistency of our study. We discovered underlying patterns, dependencies, and clusters within the dataset using principle component analysis (PCA), hierarchical clustering, and k-means clustering. These findings provided insights into the intricate interactions between the different indicators.

We were able to determine important variables and their associations by utilizing multivariate methods, which will be helpful for international organizations, and politicians. The investigation of

data structure and variability was made easier by the use of PCA, which allowed for dimensionality reduction while maintaining the necessary information. Furthermore, several clusters and groupings within the dataset could be identified using hierarchical and k-means clustering algorithms, providing a deeper knowledge of global trends and patterns.

This analysis is not without limitations, though. Limitations inherent in the data sources may still affect the quality and completeness of the dataset, even with careful collection and preparation. Furthermore, although PCA is a helpful tool for reducing dimensionality, it could mask some of the subtler details seen in the original dataset. Similar to this, although clustering techniques provide insights into how to organize data, the optimal number of groups and the choice of clustering algorithm are still subjective and may affect how easily the results may be interpreted.

References

[1] *Gapminder Dataset*, (unpublished).

[2] A. Jadhav, D. Pramod, and K. Ramanathan, *Comparison of Performance of Data Imputation Methods for Numeric Dataset*, Applied Artificial Intelligence **33**, 913 (2019).

[3] H. Abdi and L. J. Williams, *Principal Component Analysis*, WIREs Computational Statistics **2**, 433 (2010).

[4] F. Nielsen, *Hierarchical Clustering*, in (2016), pp. 195–211.

[5] J. A. Hartigan, *Statistical Theory in Clustering*, J Classif **2**, 63 (1985).

CODE:

Data loading and Preparation and Explorative data Analysis:

```
library(readxl)
excel_file <- "DATA_MV.xlsx"
df <- read_excel(excel_file)
df_clean <- df[complete.cases(df) | rowSums(is.na(df)) <= 5, ]

#converting columns
# Function to convert values
convert_k_to_numeric <- function(x) {
  # Extract numeric part and multiply by 1000
  numeric_value <- as.numeric(sub("k", "", x)) * 1000
  return(numeric_value)
}
df_clean[,6] <- sapply(df_clean[,6], convert_k_to_numeric)
df_clean[,14] <- sapply(df_clean[,14], convert_k_to_numeric)
df_clean[,-1]<- lapply(df_clean[,-1], function(x) gsub("−", "-", x))
df_clean[,-1]<- lapply(df_clean[,-1], as.numeric)
df_clean[,-1] <- as.data.frame(df_clean[,-1])

# Load the corrplot library
library(corrplot)
# Compute the correlation matrix
correlation_matrix <- cor(df_clean[,-1], use = "pairwise.complete.obs")

# Plot the correlation matrix
corrplot(
  correlation_matrix,
  method = "color",
  type = "upper",  # Display only the upper triangle of the correlation matrix
  tl.cex = 0.6,    # Adjust the size of the text labels
  diag = TRUE,    # diagonal elements
  col = colorRampPalette(c("#FFFFFF", "#7CAE00", "#2B4817"))(100),
  addCoef.col = "black",
  tl.col = "black",
  tl.srt = 45,
  number.cex = 0.5,
  mar = c(0,0,0,0),
  )
df_clean_og<-df_clean
df_clean<-df_clean[,-1]

# Function to calculate similarity between two rows
calculate_similarity <- function(row1, row2) {
  # Get indices of columns with missing values in both rows
  missing_cols_row1 <- which(is.na(row1))
  missing_cols_row2 <- which(is.na(row2))

  # Remove columns with missing values from both rows
  non_missing_indices <- setdiff(1:length(row1), c(missing_cols_row1, missing_cols_row2))
  non_missing_row1 <- row1[non_missing_indices]
  non_missing_row2 <- row2[non_missing_indices]
```

```
  # Calculate Euclidean distance between non-missing values
  euclidean_distance <- sqrt(sum((non_missing_row1 - non_missing_row2)^2))

  # Calculate Euclidean similarity (inverse of distance)
  similarity_score <- 1 / (1 + euclidean_distance)
  return(similarity_score)
}
similarity_matrix <- matrix(NA, nrow(df_clean), nrow(df_clean))

# Calculate similarity scores between each pair of rows
for (i in 1:nrow(df_clean)) {
  for (j in 1:nrow(df_clean)) {
    similarity_matrix[i, j] <- calculate_similarity(df_clean[i, ], df_clean[j, ])
  }
}
similarity_matrix

# Find top 10 most similar rows for each row
top_10_similar_rows <- lapply(1:nrow(df_clean), function(i) {
  top_10_indices <- order(similarity_matrix[i, ], decreasing = TRUE)[1:10]
  return(top_10_indices)
})
# Fill missing values in each row using similar rows
for (i in 1:nrow(df_clean)) {
  missing_indices <- which(is.na(df_clean[i, ]))  # Get indices of missing values in the current row
  for (missing_index in missing_indices) {
    similar_rows <- df_clean[top_10_similar_rows[[i]], ]  # Get similar rows
    similar_values <- similar_rows[, missing_index]  # Get values of the missing column from similar rows
    similar_values <- as.numeric(similar_values[[1]])
    df_clean[i, missing_index] <- mean(similar_values, na.rm = TRUE)  # Fill missing value with mean of simi-
lar values
  }
}
summary(df_clean)
```

Modelling:

```
df_clean_og_subset <- df_clean_og[, c("COUNTRY", "Cell phones","Children and elderly (per 100 adults)','GDP
10 yrs')]

merged_df <- merge(df_clean_og_subset, df_clean, by = c("Cell phones", "Children and elderly (per 100
adults)','GDP 10 yrs'))

numeric_cols <- merged_df[, sapply(merged_df, is.numeric)]

# Perform PCA on df_cleaned

pca_result <- prcomp(df_clean, scale. = TRUE)

# Summary of PCA

summary(pca_result)

# Scree plot

screeplot(pca_result, type = "lines")
```

```
# Create the biplot
biplot(pca_result, col = c("blue", "red"), cex = 0.7)
title(main = "Biplot of PCA Results", font.main = 1)
xlabel <- paste("PC1 (", round(summary(pca_result)$importance[2,1] * 100, 1), "%)", sep = "")
ylabel <- paste("PC2 (", round(summary(pca_result)$importance[2,2] * 100, 1), "%)", sep = "")
abline(h = 0, v = 0, col = "gray", lty = 2)
text(x = c(max(pca_result$ind$coord[,1]) * 1.1, min(pca_result$ind$coord[,1]) * 1.1),
   y = c(0, 0),
   labels = c(xlabel, ylabel),
   pos = 4, col = "black")

# Select the numeric columns for clustering (excluding non-numeric and key columns)
numeric_cols <- merged_df[, sapply(merged_df, is.numeric)]
# Perform hierarchical clustering
hc_result <- hclust(dist(numeric_cols), method = "complete")
# Plot dendrogram with country names on the x-axis
plot(hc_result, hang = -1, cex = 0.6, main = "Dendrogram of Hierarchical Clustering", labels =
merged_df$COUNTRY)

# Perform hierarchical clustering
hc_result <- hclust(dist(numeric_cols), method = "complete")
# Cut the dendrogram to obtain clusters
clusters <- cutree(hc_result, k = 5)
# Add cluster labels to merged dataframe
merged_df$Cluster <- clusters
# Print the counts of countries in each cluster
table(merged_df$Cluster)
# Plot dendrogram with clusters
plot(hc_result, hang = -1, cex = 0.6, main = "Dendrogram of Hierarchical Clustering with 5 Clusters", labels =
merged_df$COUNTRY)
rect.hclust(hc_result, k = 5, border = 2:6)  # Highlight clusters with rectangles
# Load the necessary library for clustering
library(cluster)
```

```
numeric_cols <- merged_df[, sapply(merged_df, is.numeric)]
# Perform PCA on df_cleaned
pca_result <- prcomp(numeric_cols, scale. = TRUE)
pca_result
# Extract the principal components from the PCA result
principal_components <- pca_result$x
# Choose the number of clusters
num_clusters <- 7
# Perform clustering using K-means algorithm
kmeans_result <- kmeans(principal_components, centers = num_clusters)
# Get the cluster assignments for each data point
cluster_assignments <- kmeans_result$cluster
# Print cluster centers
print(kmeans_result$centers)
# Print cluster assignments
print(cluster_assignments)
# Load the necessary library for plotting
library(ggplot2)
library(ggrepel)
# Combine principal components with cluster assignments, IDs, and country names
data <- data.frame(principal_components, cluster = as.factor(cluster_assignments), id = 1:nrow(princi-
pal_components), country = merged_df$COUNTRY)
# Plot the clusters along with IDs and country names
ggplot(data, aes(x = PC1, y = PC2, color = cluster, label = country)) +
 geom_point() +  # Plot points
   # Add labels without overlapping
 labs(title = "Cluster Plot") +  # Add title
 geom_text_repel() +
 theme_minimal()  # Set plot theme
```