Country Profiling Using Principal Component Analysis And K-Means Clustering

Patience Anipa

June 18, 2024

-

Contents

1 Introduction **1**

2 Exploratory Data Analysis **3**
2.1 Univariate Analysis 3
2.2 Bivariate Analysis 6
2.3 Multivariate Correlation Analysis 8

3 Multivariate Analysis **11**
3.1 Principal Component Analysis(PCA) 11
3.2 K - Means Clustering 15
3.3 Analysis of Results 18

4 Conclusion **23**

Chapter 1

Introduction

Motivation

The issue of efficiently optimizing the allocation of resources or aid to underdeveloped countries especially during times of crisis, disasters or natural calamities has always been one of a hurdle. In order to strategically and effectively allocate resources to the countries which require the most need in such unpredictable moments, it is essential to categorise countries in such a manner that countries that need the most help are located quickly.

The objective of this project is to address the above mentioned issue by determining the overall development of a country by clustering them according to numerical features such as social parameters, economic indicators and health factors. This will facilitate fast and easy allocation of funds to the respective countries during calamities and natural disasters. The categorization is done using Principal Components Analysis(PCA) and K-Means clustering. PCA is done to reduce dimensionality and ultimately identify the factors that contribute most to the variation in the metrics and K-Means clustering is performed to group the countries based on their respective needs.

Research Questions

The author seeks to answer, as well as provide detailed explanations to the following research questions:

- Can the variances in the metrics be explained with fewer components? What are these components and by how much can they explain the variances?
- To which extent are the different attributes correlated to each other?
- What are the necessary models or algorithms that can be used to classify countries into appropriate groups of similarity?
- Can we successfully cluster countries with similar characteristics as they are likely to face similar challenges during times of crises?

Dataset Description

The data used in this project is collected by HELP International, an international humanitarian NGO that is committed to fighting poverty and providing people of undeveloped countries with basic amenities and relief during times of disaster and natural calamities. The data can be found on Kaggle using the following link: Unsupervised Learning on Country Data (kaggle.com)

The dataset contains records of countries based on socio-economic and health factors. There are 167 rows and 10 columns in the data. Each row corresponds to a country and each column corresponds to a specific attribute of each country. The specific attributes are explained in detail in the following table.

	Attribute	Information
1	**Country**	Name of the country
2	**Child Mortality**	Death of children under 5 years of age per 1000 live births
3	**Exports**	Exports of goods and services per capita. Given as %age of the GDP per capita
4	**Health**	Total health spending per capita. Given as %age of GDP per capita
5	**Imports**	Imports of goods and services per capita. Given as %age of the GDP per capita
6	**Income**	Net annual income per person
7	**Inflation**	The measurement of the annual growth rate of the Total GDP
8	**Life Expectancy**	The average number of years a new born child would live if the current mortality patterns are to remain the same
9	**Total Fertility**	The number of children that would be born to each woman if the current age-fertility rates remain the same
10	**GDPP**	The GDP per capita. Calculated as the Total GDP divided by the total population.

Fig. 1.1: Attributes and their description.

Chapter 2

Exploratory Data Analysis

This Chapter seeks to understand the data by exploring the distribution of the data and various relationships between attributes using statistical methods and visualizations.

2.1 Univariate Analysis

To uncover an overview of how the data looks like, the following summary statistics is shown in tables 2.1 and 2.2.

	count	mean	std	min	25%	50%	75%	max
income	167.000000	17144.688623	19278.067698	609.000000	3355.000000	9960.000000	22800.000000	125000.000000
gdpp	167.000000	12964.155689	18328.704809	231.000000	1330.000000	4660.000000	14050.000000	105000.000000
child_mort	167.000000	38.270060	40.328931	2.600000	8.250000	19.300000	62.100000	208.000000
exports	167.000000	41.108976	27.412010	0.109000	23.800000	35.000000	51.350000	200.000000
imports	167.000000	46.890215	24.209589	0.065900	30.200000	43.300000	58.750000	174.000000
inflation	167.000000	7.781832	10.570704	-4.210000	1.810000	5.390000	10.750000	104.000000
life_expec	167.000000	70.555689	8.893172	32.100000	65.300000	73.100000	76.800000	82.800000
health	167.000000	6.815689	2.746837	1.810000	4.920000	6.320000	8.600000	17.900000
total_fer	167.000000	2.947964	1.513848	1.150000	1.795000	2.410000	3.880000	7.490000

Fig. 2.1: Summary statistics of numerical attributes

Feature	Skewness_type	Skewness_value
child_mort	Positively skewed	1.450774
exports	Positively skewed	2.445824
health	Positively skewed	0.705746
imports	Positively skewed	1.905276
income	Positively skewed	2.231480
inflation	Positively skewed	5.154049
life_expec	Negatively skewed	-0.970996
total_fer	Positively skewed	0.967092
gdpp	Positively skewed	2.218051

Fig. 2.2: Skewness of numerical attributes

To further analyse the distribution of the data, the histograms of all numerical attributes are plotted below.

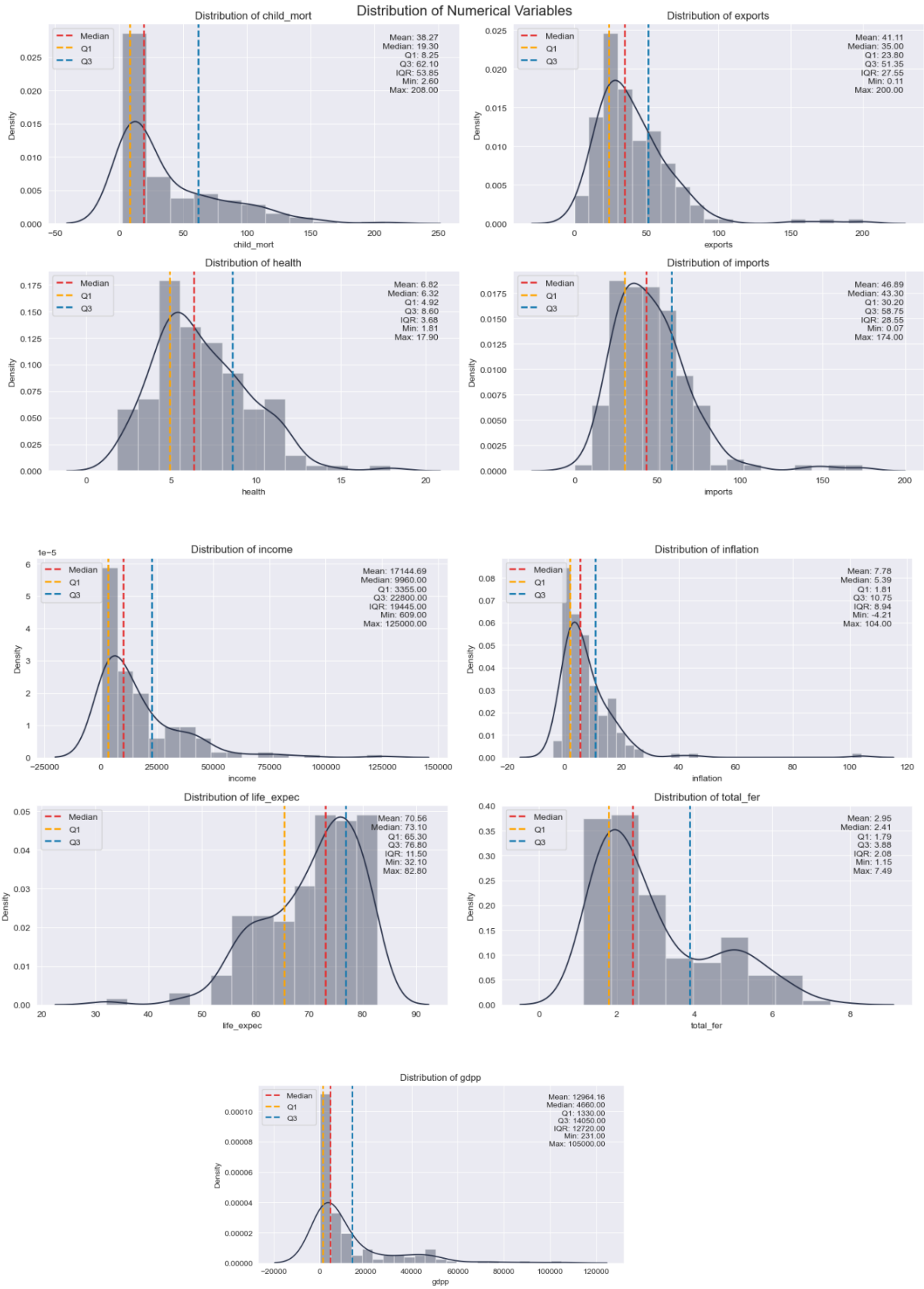

Fig. 2.3: Histograms of all numerical attributes

Observations:

- The data is characterized by highly skewed distributions, with a prominent positively skewed shape for most variables, including income, and a negatively skewed shape for only the life expectancy variable.
- The four variables, namely child mortality, exports, health, and imports, exhibit a positive skewness and have numerous outliers. Especially child mortality, which is highly skewed to the right. This means that while most countries have low child mortality rates, a few countries have very high rates. The median value is much closer to the 25th percentile than to the 75th percentile. This could be a good factor that will help in the clustering process.
- Both income per capita and inflation exhibit positive skewness, with a noticeable right-skewed distribution and a significant presence of outliers, particularly in the case of inflation. While most countries have relatively low inflation rates, there are a few countries with very high rates. Similarly, the majority of countries have an income per capita lower than approximately 25,000. This signifies that very few countries earn extremely high salaries while the majority earns lower incomes.
- Life expectancy has a distinctive left-skewed (negative skewness) distribution, which makes it susceptible to outliers. While most countries have a life expectancy of over 65, there are a few countries where life expectancy is considerably lower, around 30 or 50. This means that while the majority of the countries enjoy relatively high life expectancies, there may be marginalized countries with lower access to healthcare services or socio-economic factors, leading to shorter life spans.
- Fertility rates exhibit a distinct right-skewed distribution (positive skewness), with a considerable number of observations located at the right tail of the data. While many countries have relatively low fertility rates, around 2, a significant number of countries have higher rates of 4 and 6.
- The distribution of GDP per capita is highly right-skewed, with the majority of countries having a GDP per capita below 15,000. However, there are also several countries with a GDPP that is 2-3 times higher than this amount, indicating significant variability in economic development across countries.

2.2 Bivariate Analysis

A pairplot is used to visualize the relationships between pairs of all numerical attributes.

Fig. 2.4: Pairwise relationships between numerical variables

Observations:

- As child mortality increases, life expectancy decreases.
- As child mortality increases, total fertility also increases.
- As life expectancy increases, total fertility decreases.
- As exports increase, imports also increase.

- There are potentially two distinct clusters in terms of the relationship between income and GDPP. This pattern may be due to disparities between income and GDPP among certain countries within the dataset.
- The other pairs of variables do not seem to exhibit any form of correlation with each other.

2.3 Multivariate Correlation Analysis

Definition

Correlation is a measure that explores the linear relationship between two quantitative variables. A positive correlation signifies that one variable increases as the values of the other increases. A negative correlation on the other hand signifies that one variable decreases as the values of the other increases. Understanding the correlation between the attributes facilitates and validates further multivariate analysis.

Why Correlation Analysis?

As the objective of this project is to categorise countries based on certain features, clustering analysis is implemented in this project. The goal of clustering analysis is to group similar observations or data points into groups, and this aligns with the theme of this project. However, the quality of clustering heavily depends on the choice of variables used for clustering. Highly correlated variables can strongly impact clustering analyses, especially if the algorithm used for clustering is distance-based.

Distance-based clustering algorithms(eg. K-Means clustering) measure the distance between data points in a high-dimensional space and use this distance measure to separate similar group points into clusters. However, when variables are highly correlated, the distance measure between observations can be exaggerated, leading to biased clustering results.

To give an example, suppose we are clustering observations based on their socio-economic status, and we include both income and education as variables. These variables are often highly correlated because people with higher incomes tend to have higher education levels. If we use both income and education for clustering, the algorithm might overemphasize the similarities or differences between observations, leading to biased clustering results. In such cases it is best to drop one of the variables, unless there is a serious reason to keep them, to avoid misleading clustering results.

In summary, highly correlated variables can distort the similarities or differences between observations, leading to inaccurate clustering results. Hence, it is crucial to choose the right set of variables for clustering. Dropping correlated variables, when appropriate, can improve the quality of clustering.

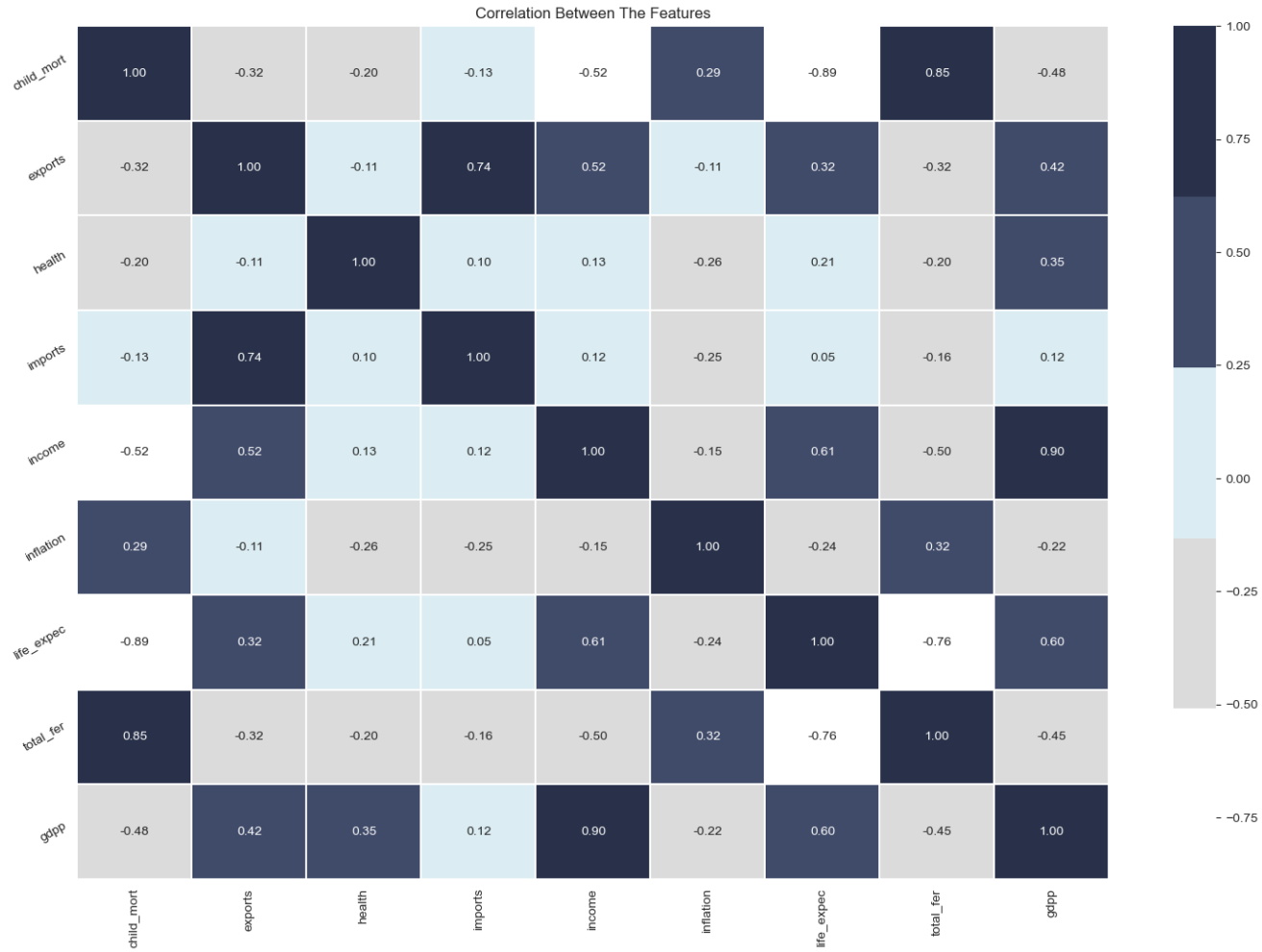

Fig. 2.5: Correlation analysis of attributes

Observations:

A coefficient greater than 0.7 is indicative of strong correlation. We explore the features that have strong correlations below.

- Child mortality and life expectancy have a strong negative correlation of 0.88.
- A strong positive correlation of 0.84 exists between child mortality and total fertility.
- Imports and exports are positively correlated with a coefficient of 0.73.
- There is a very strong positive correlation of 0.89 between income and GDP, as expected.
- Fertility and life expectancy are negatively correlated with a coefficient of 0.76.

From our observation, some of these variables are highly correlated and record very similar features. We shall drop some of them from our clustering analysis. **Principal Component Analysis (PCA)** is used to reduce the dimension in the dataset and find the features that explain most of the variation in the dataset.

Chapter 3

Multivariate Analysis

3.1 Principal Component Analysis(PCA)

Theoretical Background

Principal Component Analysis (PCA) is a technique used to reduce the dimensions of a dataset while retaining as much information as possible. This results in a dataset with fewer dimensions, but without any significant loss of information. The goal of PCA is to select the top few principal components that explain most of the variance in the data. Additionally, the new variables created through PCA are uncorrelated, which can be useful for further analysis.

PCA Analysis

The total number of principal components is 9(Variable "country" was not included since it is not a numerical variable). PCA is used to reduce the dimension of the data to fewer components to account for components which represent the most variation of the data.

A scree plot can be used to determine the optimal number of principal components for the analysis. A scree plot presents a visualization of the percentage of variance captured by each principal components as well as the cumulative sum of the explained variances which can be used to determine how many components to select for a given level of retained variance.

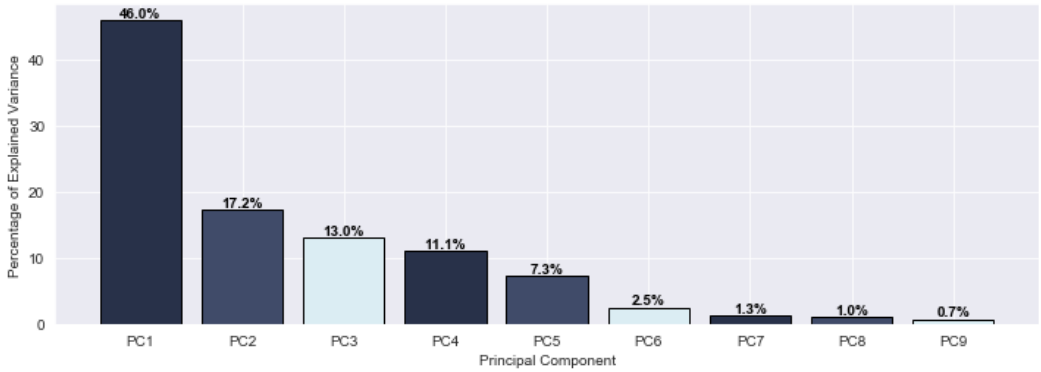

Fig. 3.1: Scree plot of PC1 - PC9 explained variance.

Observations:

- The first principal component explains the most variance, that is approximately 46% of the total variance in the data, while the later components explain progressively less.

- By checking the cumulative proportion, PC1 to PC5 explain about 94.6% of the total variance.

The range of variance that is generally considered to be a good choice for choosing the optimum number of components in PCA is between 70% and 95%. According to the PCA analysis, we can keep around 94.6% of the information from our data with only 5 dimensions. Thus, we reduce the number of features on our dataset from 9 to 5 numeric features.

The following figures shows plots of the PCA loadings and scores.

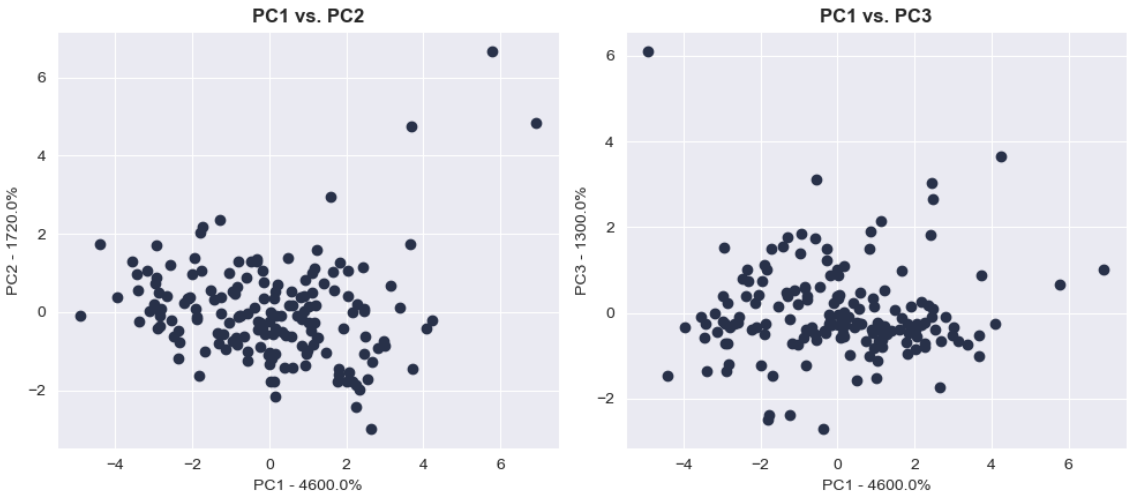

Fig. 3.2: PCA scores plot

Fig. 3.3: Loadings plot

PCA Biplots

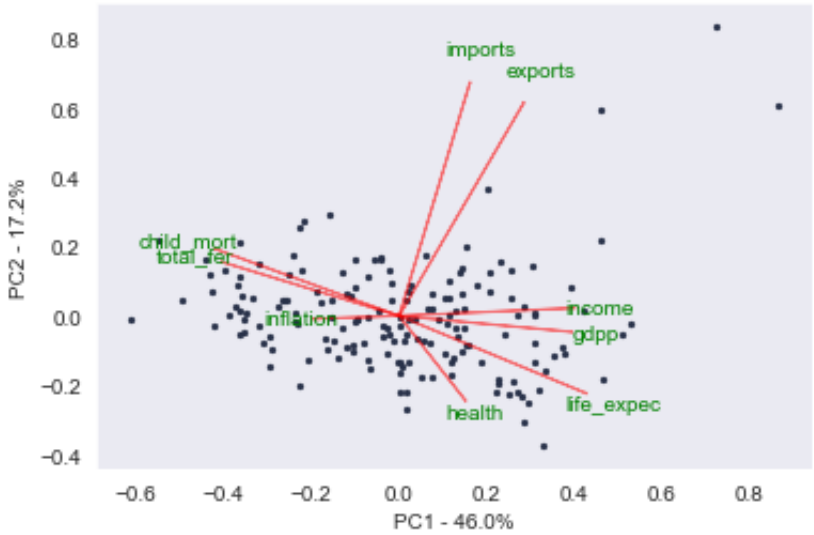

Fig. 3.4: Biplot of PC1 and PC2

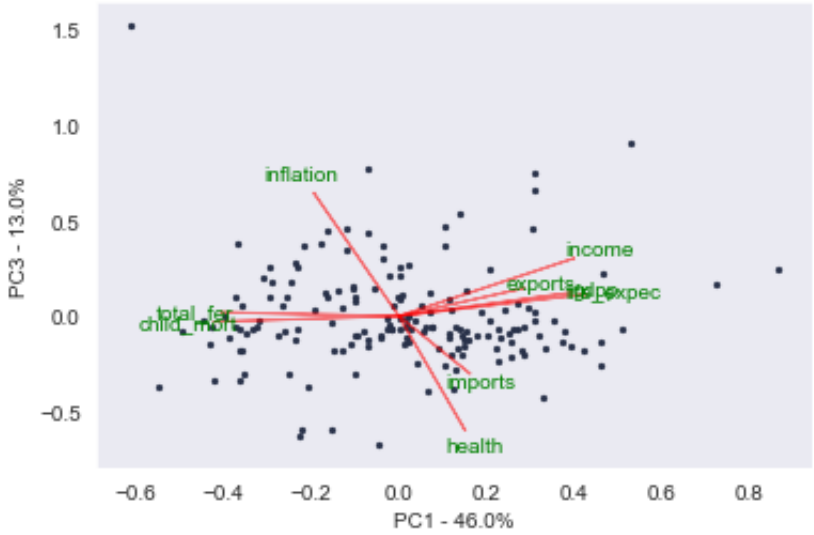

Fig. 3.5: Biplot of PC1 and PC3

Observations

- Generally, PC1 has positive associations with exports, health, imports, income, life expectancy, and gdpp, while it has negative associations with child mortality, inflation, and total fertility. This means that countries with high values for the first principal component are generally stable, as they have high ratings for exports, imports, income, health, life expectancy, and gdpp, but low ratings for child mortality, inflation and total fertility.
- The Biplot of PC1 and PC2 capture most of the variance in the data. The angles between the vectors show that the following variables are positively correlated: imports and exports, child mortality and total fertility, income and GDP. Also note

that since the variables child mortality and life expectancy diverge and form an angle close to 180°, they are negatively correlated.

  - Imports and exports have the most influence since they are farthest away from the origin. It can be seen that the countries affected by these factors are almost insignificant.
  - On the top left axis, child mortality and total fertility have positive values showing that those countries are dominated by high rates of child mortality and fertility.

- The Biplot of PC1 and PC3 possess similar characters to that of the first two principal components. However, inflation has positive values while child mortality and inflation lie very close to the origin of PC1 and PC3. Also, exports and imports have low correlation.

3.2 K - Means Clustering

Theoretical Background

The basic idea of K-means clustering is to group a number of observations(n) into a number of clusters(K) based on their distance from one another. In implementing the algorithm the goal is to minimize the sum of squared distances between the data points and their corresponding cluster centroids, resulting in clusters that are internally homogeneous and distinct from each other.

The Hartigan-Wong Algorithm

The Hartigan-Wong algorithm(Hartigan and Wong 1979) is implemented here. This algorithm searches for the partition of data space with locally optimal within-cluster sum of squares of errors(SSE). It may assign a case to another subspace, even if it currently belongs to the subspace of the closest centroid, if doing so minimizes the total within-cluster sum of square. Summarily, the process of the algorithm begins by first initializing the clusters centres, then cases are then assigned to the centroid nearest them and the centroids are calculated as the mean of the assigned data points. The iterations are as follows. If the centroid has been updated in the last step, for each data point included, the within-cluster sum of squares for each data point if included in another cluster is calculated. If one of the cluster sum of square (SSE2 in the equation below, for all $i \neq 1$) is smaller than the current one (SSE1), the case is assigned to this new cluster.

$$SSE2 = \frac{N_i \sum_j \|x_{ij} - c_i\|^2}{N_i - 1} < SSE1 = \frac{N_1 \sum_j \|x_{1j} - c_1\|^2}{N_1 - 1}$$

The iterations continue until a change would make the clusters more internally variable or more externally similar.

The sum of squared Euclidean distances between elements and the corresponding centroid is given below:

$W(C_k) = \sum_{x_i \in C_k} (x_i - \mu_k)^2$

x_i designs a data point belonging to the cluster C_k

μ_k is the mean value of the points assigned to the cluster C_k

The total within-cluster variation is defined below:

$\sum_{k=1}^{k} W(C_k) = \sum_{k=1}^{k} \sum_{x_i \in C_k} (x_i - \mu_k)^2$

The algorithm is implemented as follows:

- To start the algorithm, k centroids are chosen randomly, where k is the number of

desired clusters. The distances between each data point and the centroids are then calculated, and each data point is assigned to the nearest centroid.

- Once all data points have been assigned to a centroid, the average value of the data points within each cluster is calculated, and these average values are defined as the new centroids. The process repeats itself until both centroids converge to fixed points, at which point the algorithm terminates.

Here is the pseudocode describing the iterations:
1- Choose the number of clusters
2- Choose the metric to use
3- Choose the method to pick initial centroids
4- Assign initial centroids
5- Assign cases to closest centroid
6- Calculate centroids
7- For $j <= nb$ clusters, if centroid j was updated last iteration:
a. Calculate SSE within cluster
b. For $i <= nb$ cases in cluster
i. Compute SSE for cluster $k! = j$ if case included
ii. If SSE cluster $k < SSE$ cluster j, case change cluster

Choosing the Number of Clusters(Elbow Method)

Before starting the algorithm, the number of clusters must be set. The **Elbow Method** is used to determine the optimal number of clusters for this clustering. The Elbow method uses several different values of k and examines the differences in the results to find the optimal number of clusters. In other words, it involves plotting the within-cluster sum of squares against the number of clusters and identifying the "elbow" point where the rate of decrease in the sum of squares slows down.

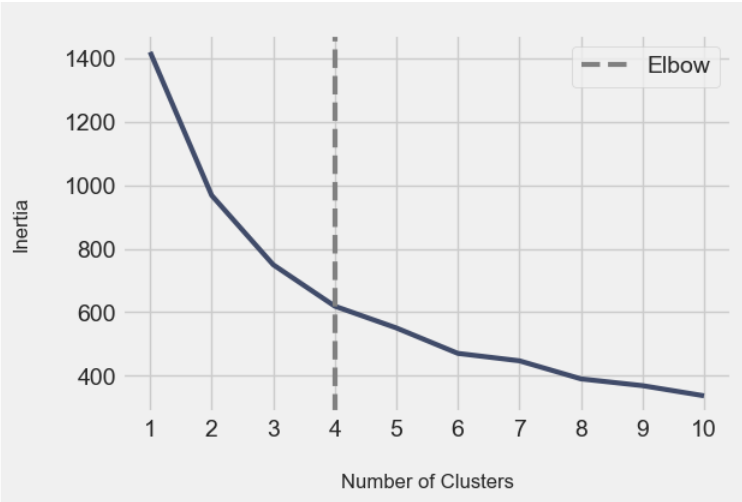

Fig. 3.6: Inertia/Within-group sum of squares vs Number of clusters.

The results suggest that the optimal number of clusters is 4 as it appears to bend the elbow at that point.

Implementing a clustering model with 4 clusters means that the dataset will be divided into four distinct groups. Each group will contain data points that are similar to each other but different from the other groups. The clustering algorithm works by iteratively comparing the similarity of each data point to the other data points in the dataset, and assigning it to the appropriate cluster based on the similarity criteria.

Getting the Coordinates for the Centroids

	PC1	PC2	PC3	PC4	PC5
0	2.585905	-0.866690	0.075067	0.982564	-0.260166
1	5.460225	5.432473	0.211648	0.906058	0.455229
2	-2.434620	0.411276	-0.096167	0.691710	-0.141458
3	0.235279	-0.110651	0.018769	-0.743741	0.150435

Fig. 3.7: Coordinates of centroids for the four clusters.

K-Means Clustering

	PC1	PC2	PC3	PC4	PC5	k-means Cluster
0	-2.913025	0.095621	-0.718118	1.005255	-0.158310	2
1	0.429911	-0.588156	-0.333486	-1.161059	0.174677	3
2	-0.285225	-0.455174	1.221505	-0.868115	0.156475	3
3	-2.932423	1.695555	1.525044	0.839625	-0.273209	2
4	1.033576	0.136659	-0.225721	-0.847063	-0.193007	3
...	...	...	...	...	...	...
162	-0.820631	0.639570	-0.389923	-0.706595	-0.395748	3
163	-0.551036	-1.233886	3.101350	-0.115311	2.082581	3
164	0.498524	1.390744	-0.238526	-1.074098	1.176081	3
165	-1.887451	-0.109453	1.109752	0.056257	0.618365	2
166	-2.864064	0.485998	0.223167	0.816364	-0.274068	2

167 rows × 6 columns

Fig. 3.8: Clustered dataset

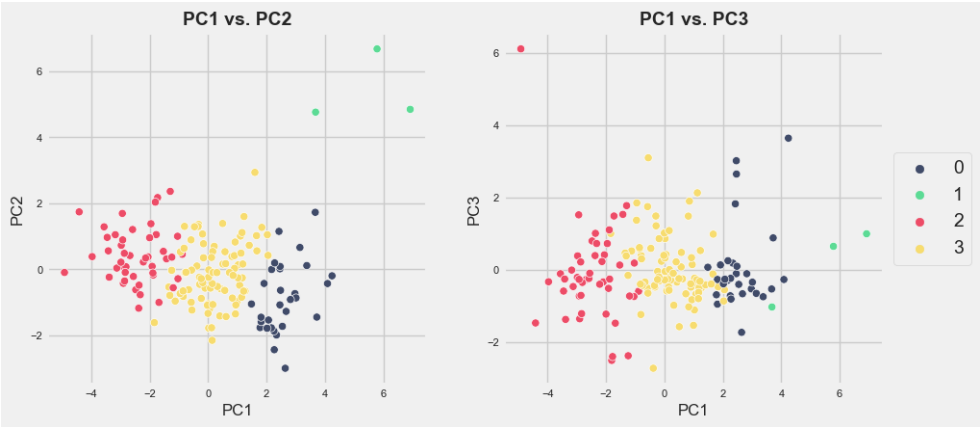

Fig. 3.9: Clusters Plot

The results from the K-Means clustering is further analysed in the following sections.

3.3 Analysis of Results

Distribution of Data by Clusters

The countries are grouped into clusters as follows.

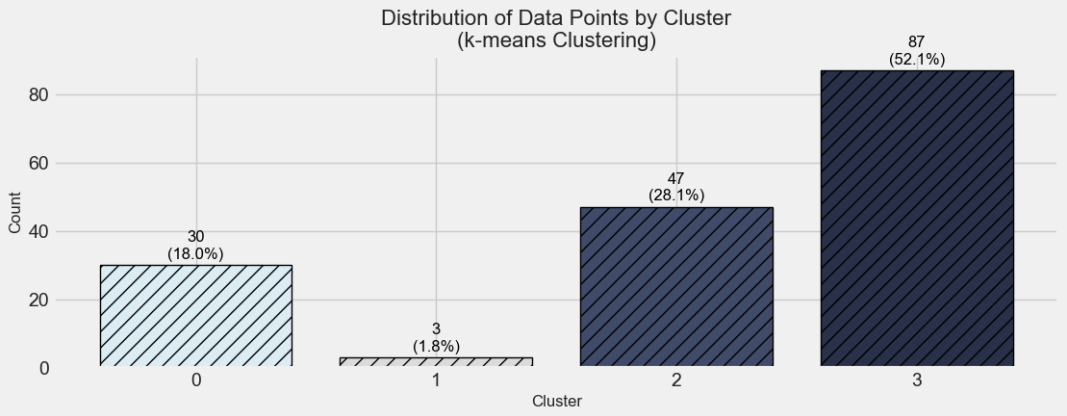

Fig. 3.10: Clustered dataset

The countries can be divided into four clusters based on certain characteristics.

- The largest cluster (3), with 87 countries, covers over 50% of the total number of countries.
- The second largest cluster (2), with 47 countries, covers approximately 28% of the total.
- The third cluster largest (0) covers 18% of the overall countries, comprising 30 countries.
- The smallest cluster (1) consists of only 3 countries, representing 1.8% of the total count.

Country Profiling

If one or more clusters have a higher concentration of countries with lower income and higher child mortality, then we can conclude that these clusters represent economically disadvantaged countries. By contrast, if we see that some clusters have a lower concentration of such countries, then we can infer that those clusters may represent more economically stable countries.

By examining the boxplots of Income or GDP, Life expectancy, Total fertility, and Child mortality across the different clusters, we can gain insights into the economic status of the countries in our dataset and how they are grouped together.

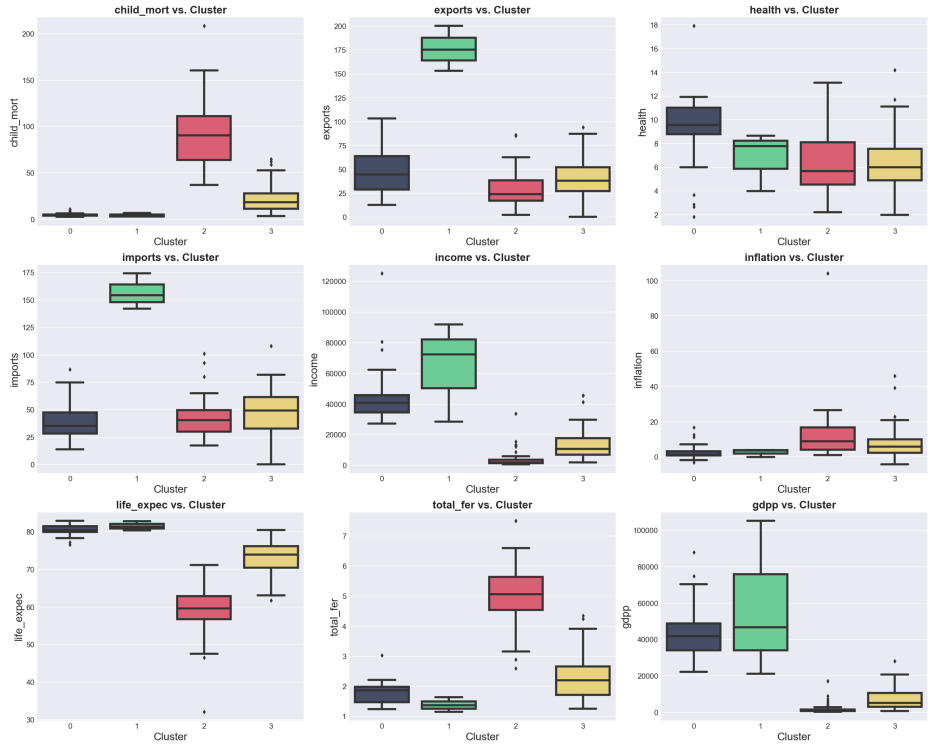

Fig. 3.11: Box plots of Features vs Clusters

Countries are profiled based on the following characteristics respectively.

- **Cluster 1 : (30 Countries) –> Developed Countries –> No Need for Financial Aid(Need in very few aspects)**
  - low 'Child mortality', high 'Income', high 'Life expectancy', and low 'Total fertility'
- **Cluster 2 : (3 Countries) –> Highly Developed Countries –> No Need for Financial Aid**
  - low 'Child mortality', high 'Income', high 'Life expectancy', and low 'Total fertility'
- **Cluster 3 : (47 Countries) –> Undeveloped Countries –> Need Financial Aid**
  - high 'Child mortality', low 'Income', moderate 'Life expectancy', and high 'Total fertility'
- **Cluster 4 : (87 Countries) –> Developing Countries –> Might Need Financial Aid**
  - moderate 'Child mortality', low 'Income', moderate 'Life expectancy', and moderate 'Total fertility'

A map is generated to visualize the clusters.

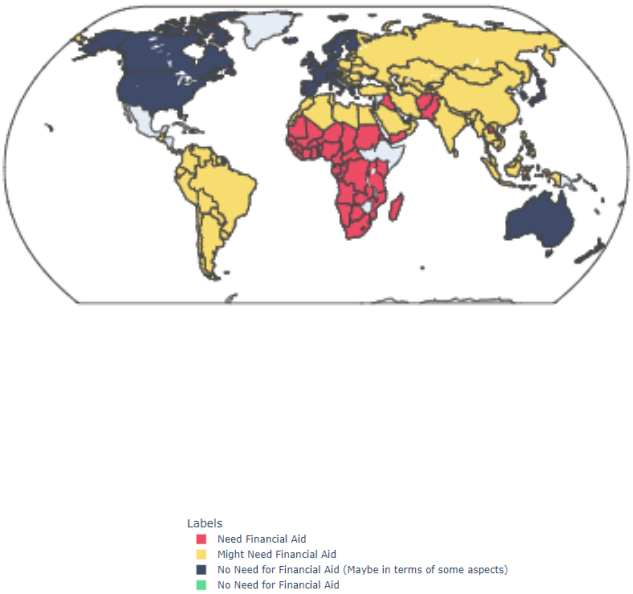

Fig. 3.12: Map showing profiled countries

The objective of this research was to easily and quickly identify countries which need urgent support in times of crises. This was achieved by categorizing countries based on their level of development which influences their urgency for need. The following categorizations are given to the countries:

Cluster 1: Developed Countries (30 countries)

(No Need for Financial Aid(Need in very few aspects))

From the analysis it was observed that the countries in this class have low child mortality rates, high income and GDP levels, high life expectancies, and low total fertility rates.These countries (Australia, North America and most European countries) are highly developed and have a very high standard of living. They have a highly diversified and advanced economy, excellent healthcare systems, and a strong focus on education. They have a strong infrastructure and are mostly self-sufficient, but may still require financial aid in some aspects, such as disaster relief or refugee support.

Cluster 2: Highly Developed Countries (3 countries)

(No Need for Financial Aid)

The countries in this class have low child mortality rates, high income and GDP levels, high life expectancies, and low total fertility rates. These countries (Luxembourg, Malta, and Singapore) are considered the most developed in the world. They have the highest standard of living and the highest level of development across all areas, including

education, healthcare, and economy. They have a strong infrastructure and are largely self-sufficient, with no need for financial aid.

Cluster 3: Undeveloped Countries (47 countries)

(Need Financial Aid)

The countries in this class have high child mortality rates, low income and GDP levels, middle life expectancies, and high total fertility rates. These countries (Most African countries) are the least developed and have the lowest standard of living. They have a predominantly agricultural economy and lack a strong infrastructure, with limited access to education and healthcare. They are heavily reliant on foreign aid and assistance to support their basic needs, such as food, water, and shelter.

Cluster 4: Developing Countries (87 countries)

(Might Need Financial Aid)

The countries in this class are characterized by moderate child mortality rates, moderate low income and GDP levels, moderate life expectancies, and moderate total fertility rates. These countries (Most Asian and South American countries) are transitioning from being underdeveloped to developed. They have made significant progress in areas such as education, healthcare, and economy, but still have a long way to go. They have a partially developed infrastructure and may require financial aid to support their growth and development. Depending on their progress, they may require financial aid for specific aspects of development or emergencies.

Chapter 4

Conclusion

This chapter provides a summary of the findings and conclusions that can be drawn from the observations.

Based on the analysis, the following address the research questions raised at the beginning of the analysis.

- The variances could be explained with fewer components. Principal Component Analysis(PCA) showed that by using the first 5 principal components, we obtained about 94.6% of the total variance of the data. The components are used for further clustering analysis.
- From the correlation matrix, the correlations between the variables were found. The cut-off value for the significant dependency between variables above 7.0 was employed for further analysis.
- The K-means clustering algorithm is used to classify similar countries into corresponding groups. The Elbow method is used to determine the optimal number of clusters and the choice of four clusters helps to classify and draws the distinct difference between the groups.
- Countries were successfully categorized/clustered into four groups, namely Developed Countries(No Need for Financial Aid(Need in very few aspects)), Highly Developed Countries(No Need for Financial Aid), Undeveloped Countries(Need Financial Aid) and Developing Countries(Might Need Financial Aid).

Critical Evaluation

The analysis is performed based on knowledge about different multivariate statistical analysis methods. During the analysis, multiple online sources were consulted to help diagnose certain formulations and the explanation of coding results.

However, the analysis explanation is written by the author's personal opinion and interpretation. For instance, classifying clusters based on the parameters, income and child mortality could not be very efficient in grouping clusters leading to a possible bias.

Additionally, different clustering algorithms propose varied number of clusters, and even though they collectively possess similar characteristics to the results of the K-Means clustering algorithm, they could be better classifications for the research purpose.

Also, since secondary data was employed in the analysis, the quality of the data could be questioned and thus affect the final results of the analysis. As mentioned above, it is likely that the results of these findings is different from others, since the choice of the parameter might be different and thus affects the final result.

It is highly recommended that one tries with different parameters to tune the accuracy of the used methods, and to see if more findings could be found and explained.