Clustering and PCA

Assignment submitted by

Janarthanan B

jnvdasa@gmail.com

Introduction

In this section, we provide a synopsis of the problem statement and the solution approach we take to address the problem.

Problem Statement

Client: HELP International (an international humanitarian NGO committed to fighting poverty and providing the people of backward countries with basic amenities and relief during the time of disasters and natural calamities)

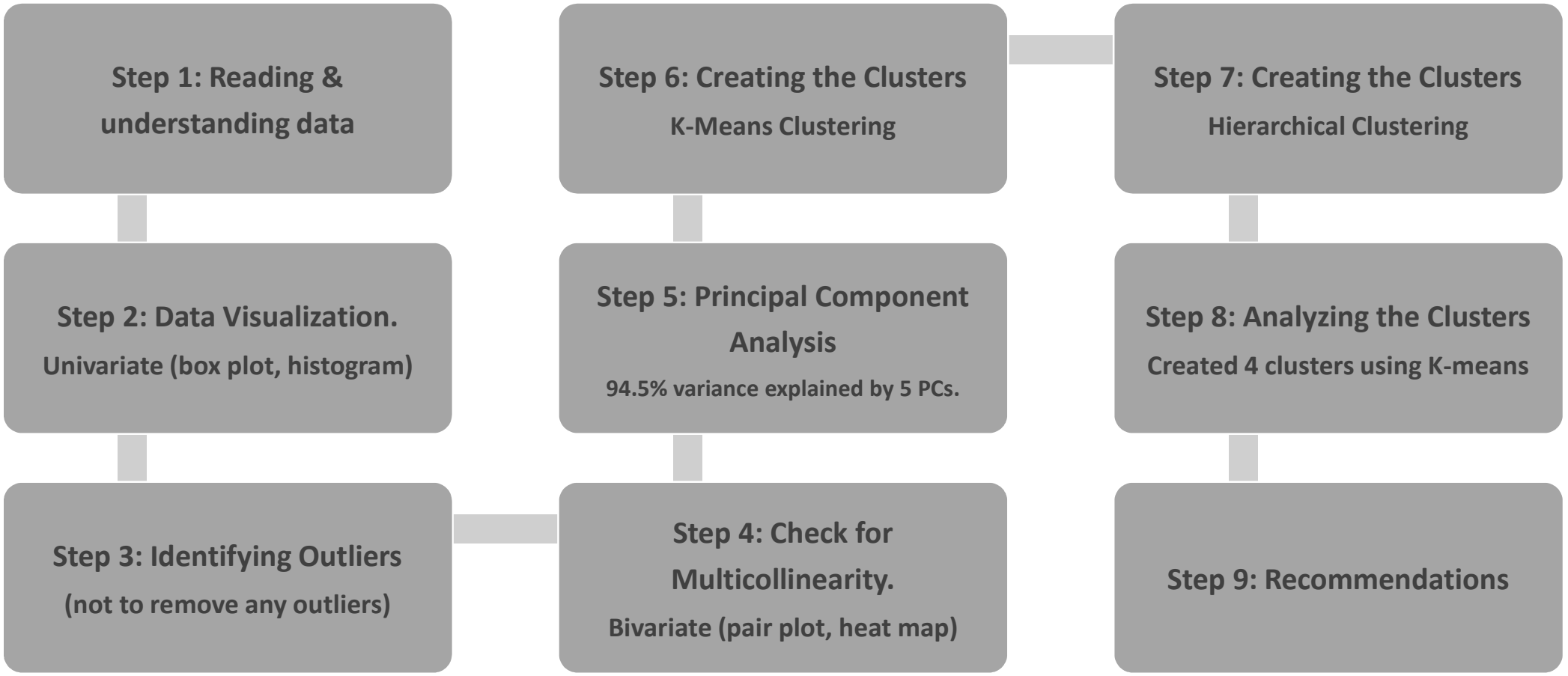
Business Question: How to use the $ 10 million funds raised effectively by choosing countries that are in the direst need of aid?

Dataset Available: Country data with 9different parameters: child mortality rate, net income per person, GDP per capita, inflation rate, life expectancy, fertility rate, import, export, spend on health, all of which are numeric variables (refer next slide for Data Dictionary)

Data Dictionary

Column Name	Description	Type	Null
country	Name of the country	String	None
child_mort	Death of children under 5 years of age per 1000 live births	Float	None
exports	Exports of goods and services. Given as %age of the Total GDP	Float	None
health	Total health spending as %age of Total GDP	Float	None
imports	Imports of goods and services. Given as %age of the Total GDP	Float	None
Income	Net income per person	Integer	None
Inflation	The measurement of the annual growth rate of the Total GDP	Float	None
life_expec	The average number of years a new born child would live if the current mortality patterns are to remain the same	Float	None
total_fer	The number of children that would be born to each woman if the current age-fertility rates remain the same.	Float	None
gdpp	The GDP per capita. Calculated as the Total GDP divided by the total population.	Integer	None

Solution Methodology



Exploratory Data Analysis

In this section, we explore the data with an objective to understand the data, address data quality issues (if any), visualize the data to get more insights, identify outliers or scaling issues in the data and to see how the variables are correlated with each other.

Reading and Understanding the Data

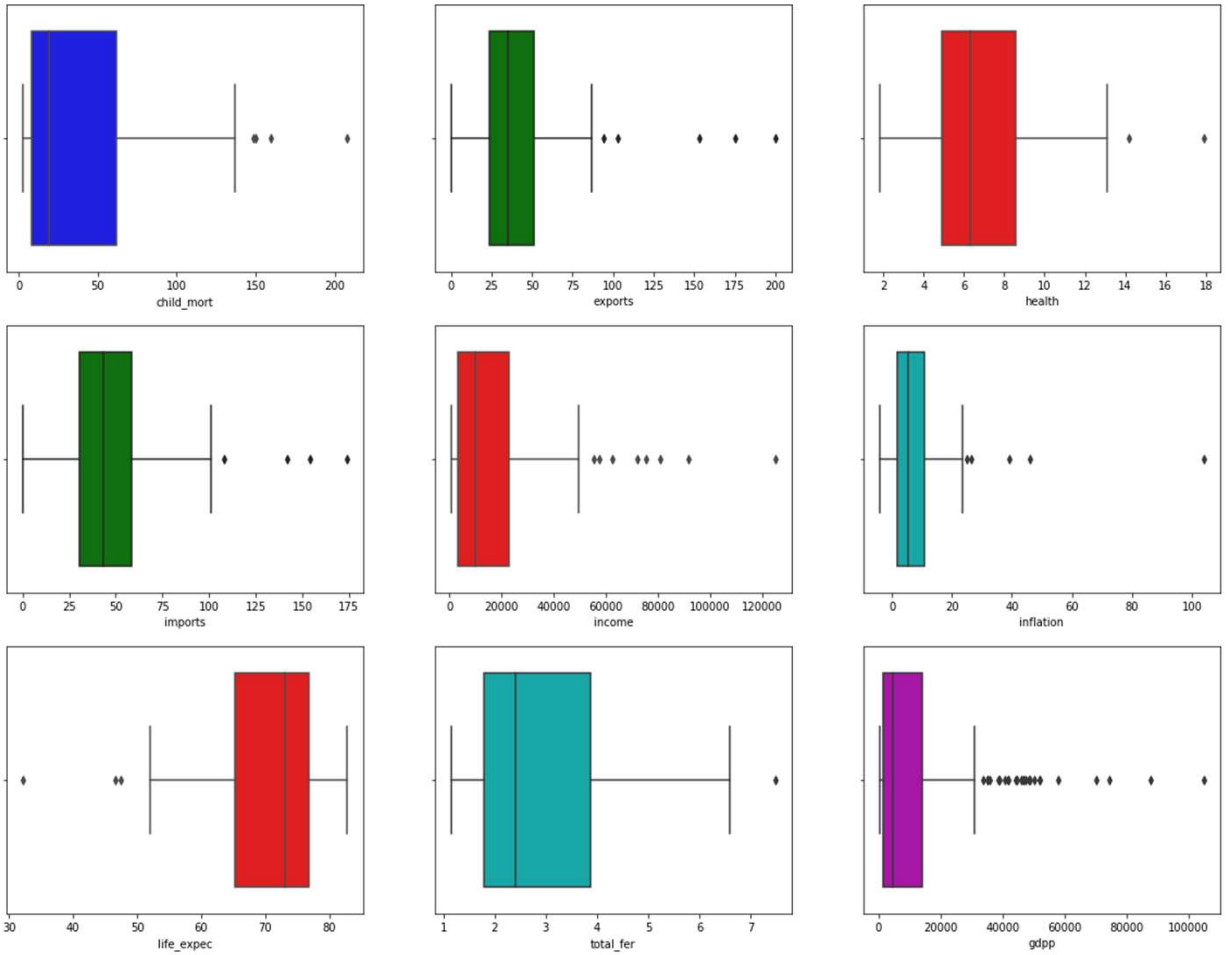
- Read the data from the csv file and load it into Python Data frame
- Check for any data quality issues and address them.
- Identify the outliers. Decide on how to treat the outliers.
- Univariate Analysis to get insight into each variable & its distribution.
- Bivariate Analysis to study the correlations between variables.

Data Quality Check

S.No.	Description	Yes	No
1	Any unnecessary rows: blank rows, header, footer, summary, total, subtotal?		✓
2	Any issues with columns: missing or inconsistent column name, unwanted column etc.		✓
3	Are there any missing values or null values?		✓
	Note: If there are missing values, we have two options: drop the row or column or impute the missing values (try other data sources, or derive the values or use business judgment)		
4	Are there any issues with the data types of the columns?		✓
5	Is there a need to standardize precision for better presentation of data?		✓
6	Are there any outliers? Is there a need to remain outliers? (Refer next slide)		✓
7	Are the observations under each variable have any variation in scale?		✓

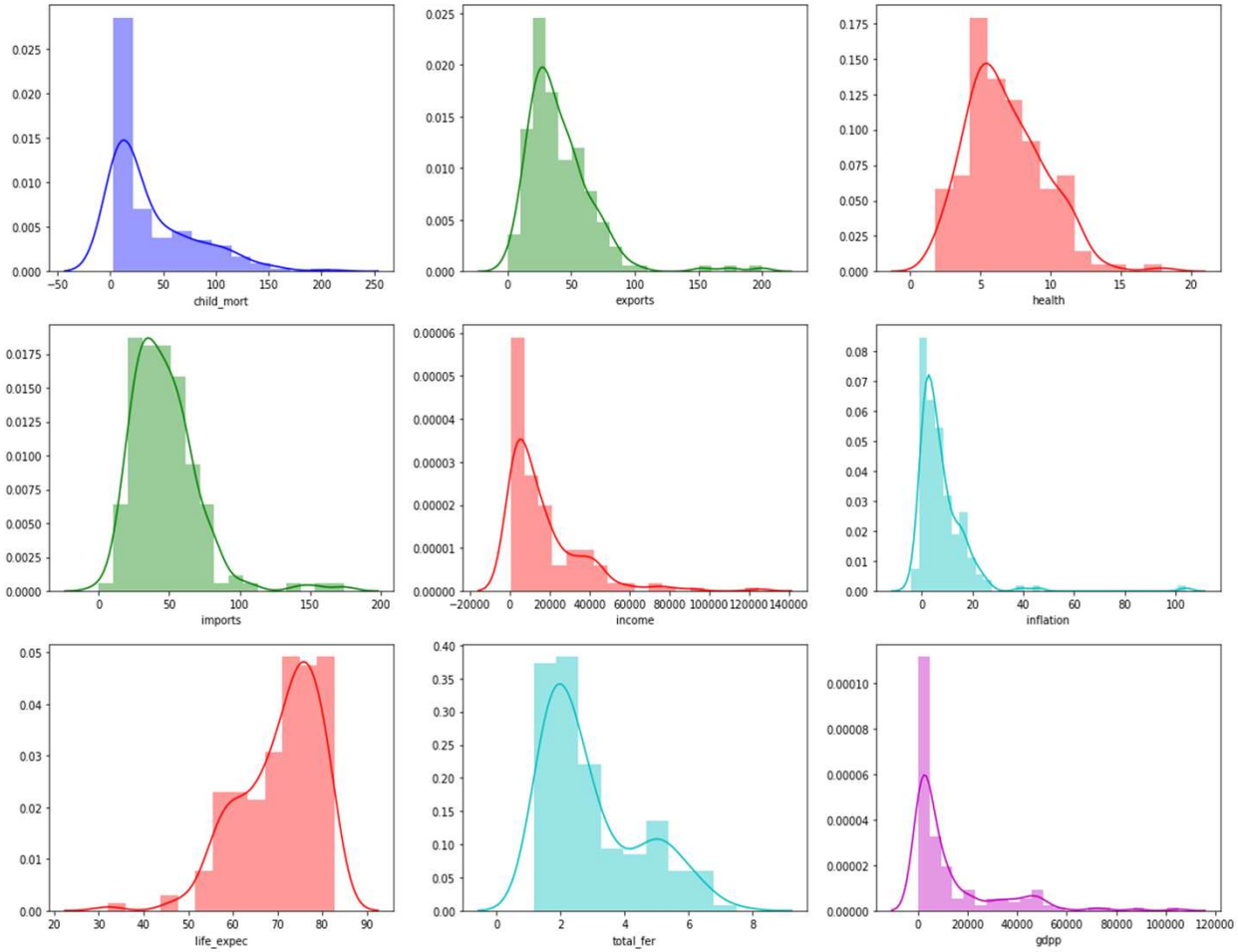
Univariate Analysis

Box and Whiskers plots are ideal for studying the data distribution and checking for any outliers. The dots beyond the whiskers indicate the outliers. There are a quite a lot of outliers in the variables income and gdpp. The line in the middle of the box is an indicator of any skewness in the data.



Univariate Analysis - contd.

Histogram and Distribution curves can also be used for analysing the individual numeric variables in the data to study the distribution and skewness. We do not get much information about the outliers here.



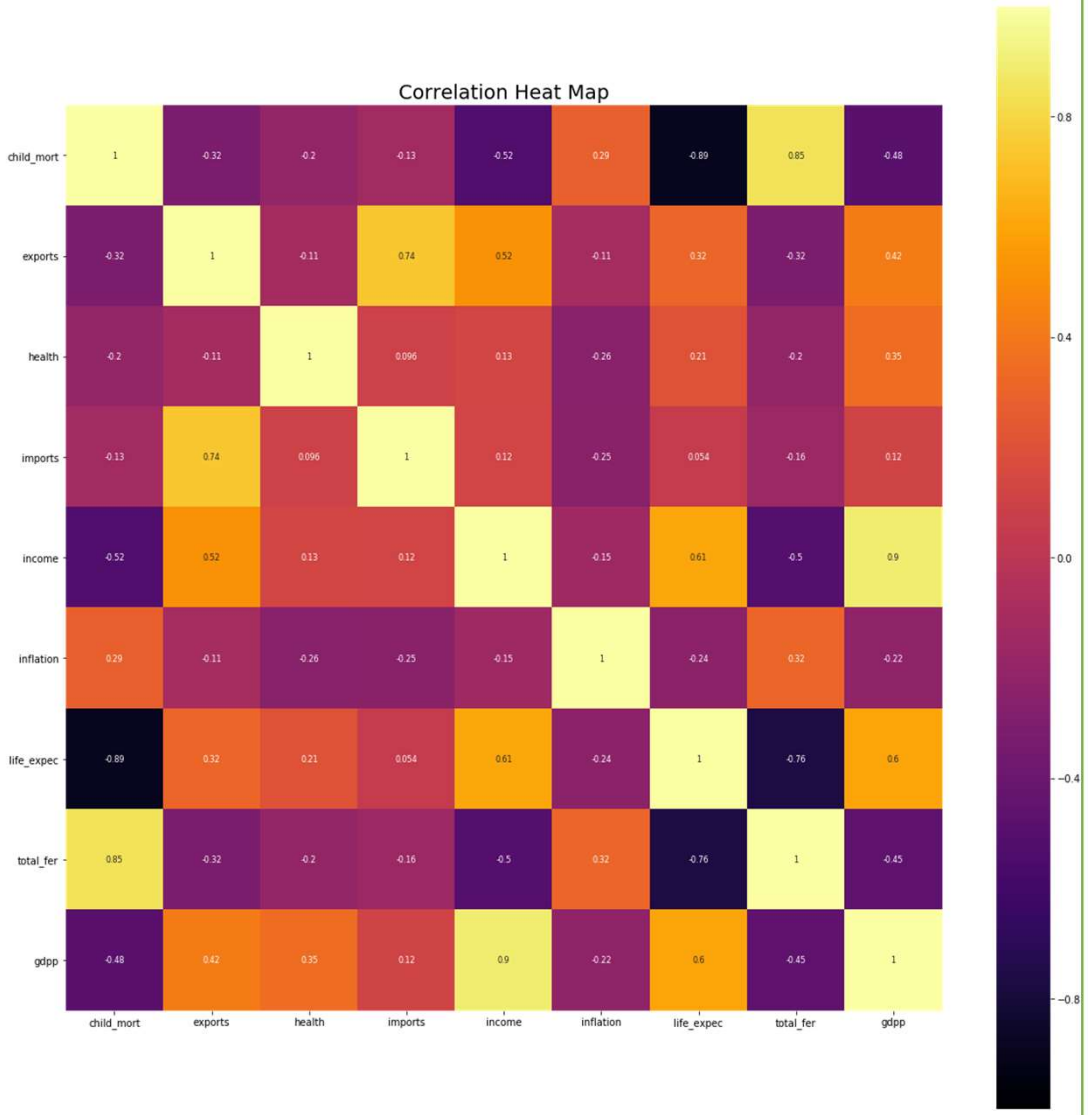
Identifying Outliers

	count	mean	std	min	25%	50%	75%	max	IQR	whisker1	whisker2	Outliers	Skew
child_mort	167.0	38.270060	40.328931	2.6000	8.250	19.30	62.10	208.00	53.850	-72.53	142.88	4.0	Right
exports	167.0	41.108976	27.412010	0.1090	23.800	35.00	51.35	200.00	27.550	-17.52	92.67	5.0	Right
health	167.0	6.815689	2.746837	1.8100	4.920	6.32	8.60	17.90	3.680	-0.60	14.12	2.0	None
imports	167.0	46.890215	24.209589	0.0659	30.200	43.30	58.75	174.00	28.550	-12.63	101.58	4.0	Right
income	167.0	17144.688623	19278.067698	609.0000	3355.000	9960.00	22800.00	125000.00	19445.000	-25812.50	51967.50	8.0	Right
inflation	167.0	7.781832	10.570704	-4.2100	1.810	5.39	10.75	104.00	8.940	-11.60	24.16	5.0	Right
life_expec	167.0	70.555689	8.893172	32.1000	65.300	73.10	76.80	82.80	11.500	48.05	94.05	3.0	Left
total_fer	167.0	2.947964	1.513848	1.1500	1.795	2.41	3.88	7.49	2.085	-1.33	7.01	1.0	Right
gdpp	167.0	12964.155689	18328.704809	231.0000	1330.000	4660.00	14050.00	105000.00	12720.000	-17750.00	33130.00	25.0	Right

There are not many outliers, except for the variables income and gdpp. However, removing the outliers may result in some countries missing out the chance of getting the funds. So we have decided not to remove any outliers from the data.

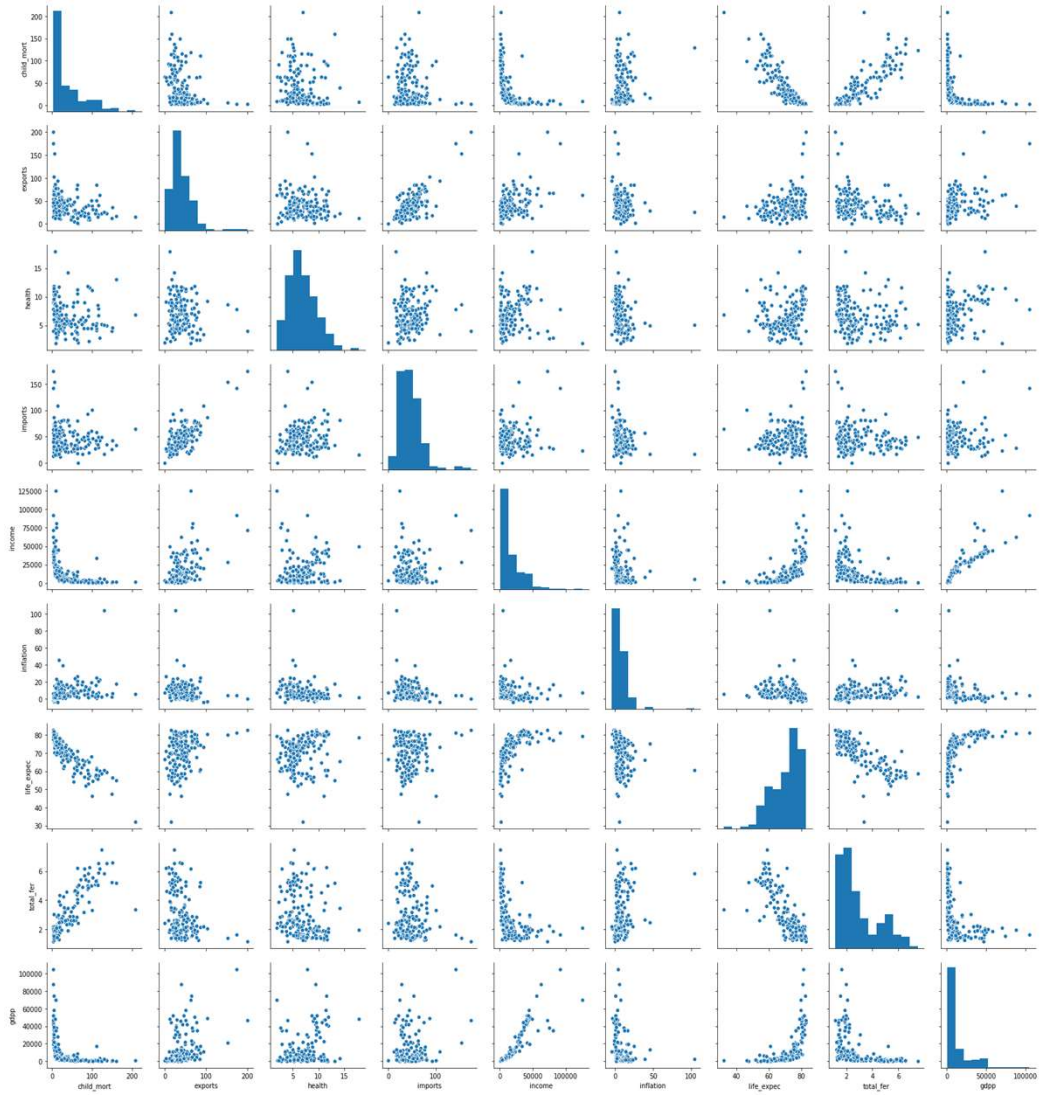
Bivariate Analysis

We see that many of the variables are strongly correlated. (1) income and gdpp has strong positive correlation; (2) child mortality and fertility rate has strong positive correlation (3) both child mortality and fertility rate has strong negative correlation with life expectancy and (4) imports and exports has strong positive correlation.



Bivariate Analysis

Multicollinearity refers to a situation in which two or more independent variables are highly linearly related. From the heat map in previous slide and the scatter plot it is evident that we have variables that are correlated with each other. With PCA (principal component analysis), we can avoid the problem of multicollinearity. The principal components are linear combinations of the original variables in such a way as to explain the variance without losing information (i.e. without the need for dropping any columns)



Principal Component Analysis

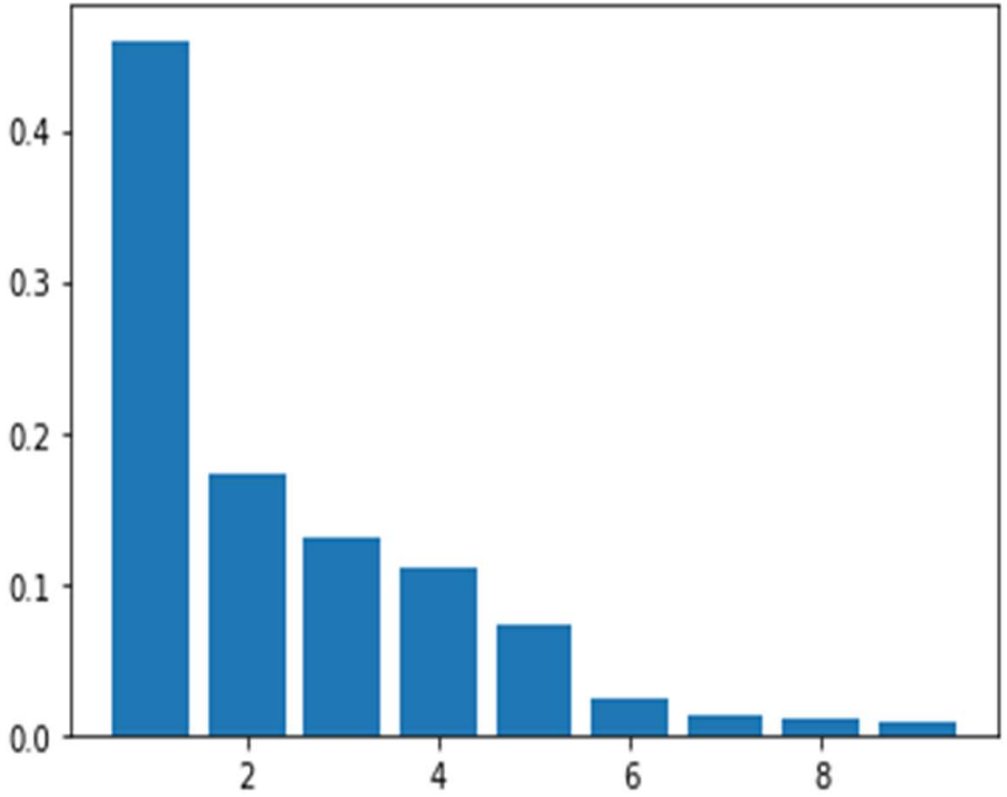
In this section, we proceed with the PCA for dimensionality reduction to arrive at a fewer non correlated variables that are linear combination of the original variables and preserve maximum information.

Standardization of the Variables

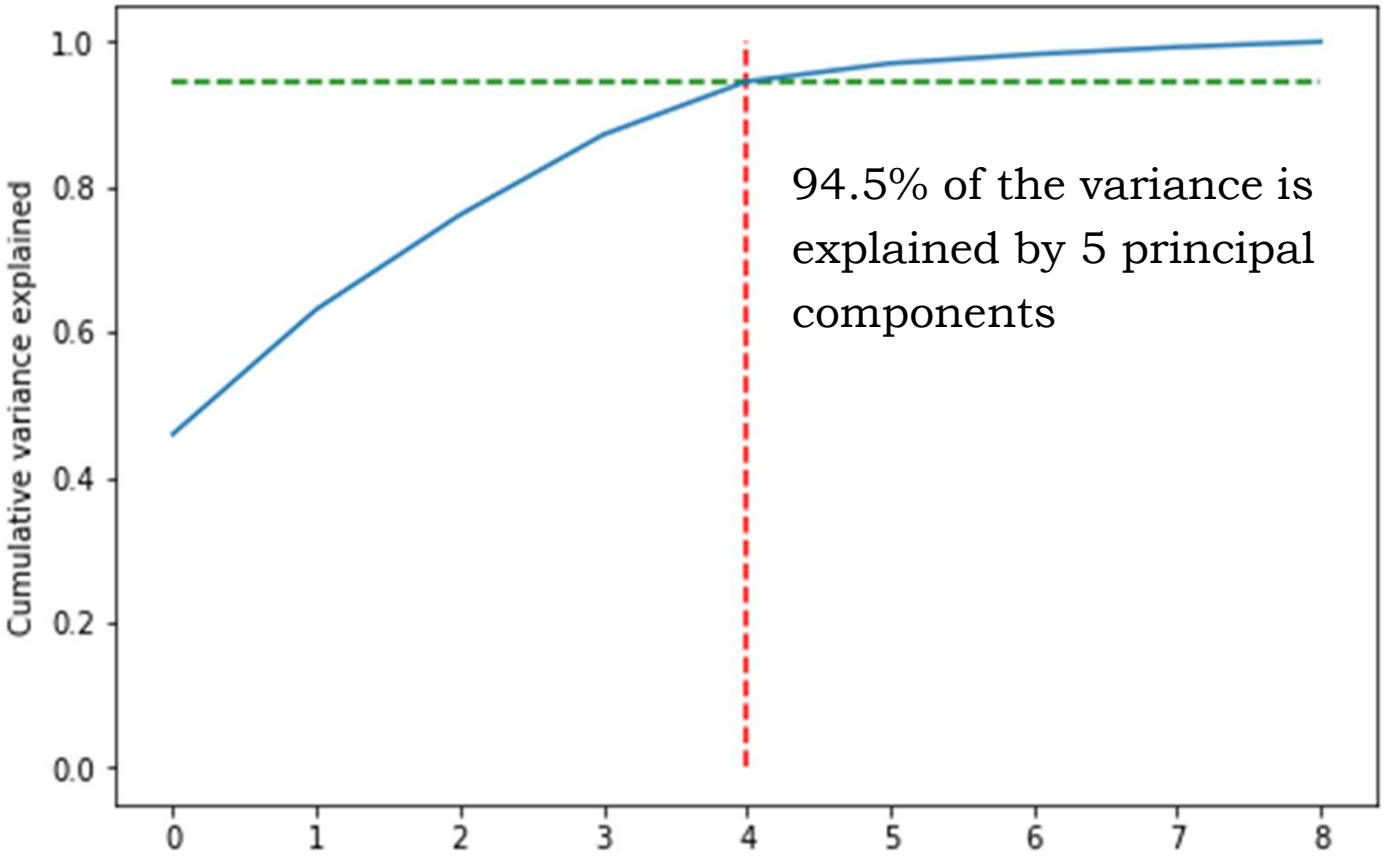
- The 9 numeric variables are in different scales.
- It is important to standardize them before performing PCA.
- Standardization is a method which scales the data in such a way that the mean = 0 and the standard deviation = 1.
- We do this using StandardScaler imported from sklearn.preprocessing

PCA explained variance ratio

If we apply PCA on n variables it creates a maximum of n variables and each principal component explains certain variance in the data. Here we have 9 principal components out of which we see that 5 of them explain almost 95% of the variance in the data.

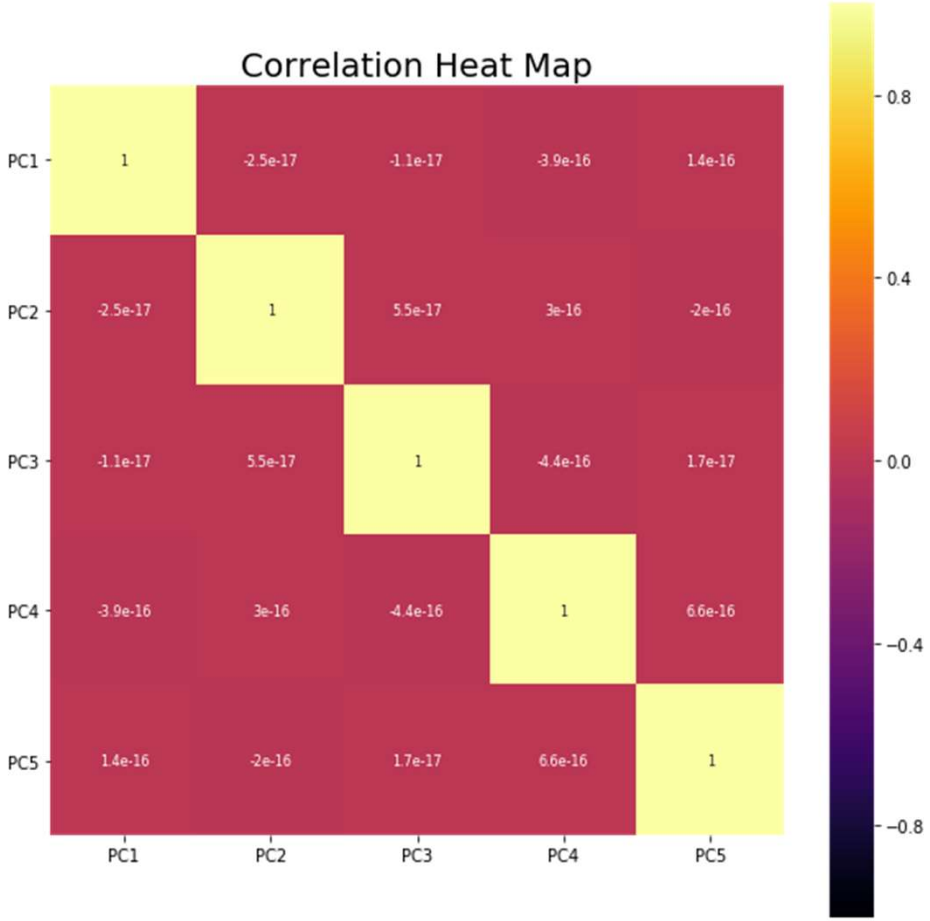


Scree Plot



Note: a scree plot is a line plot of the eigenvalues of principal components in an analysis.

Non correlated components



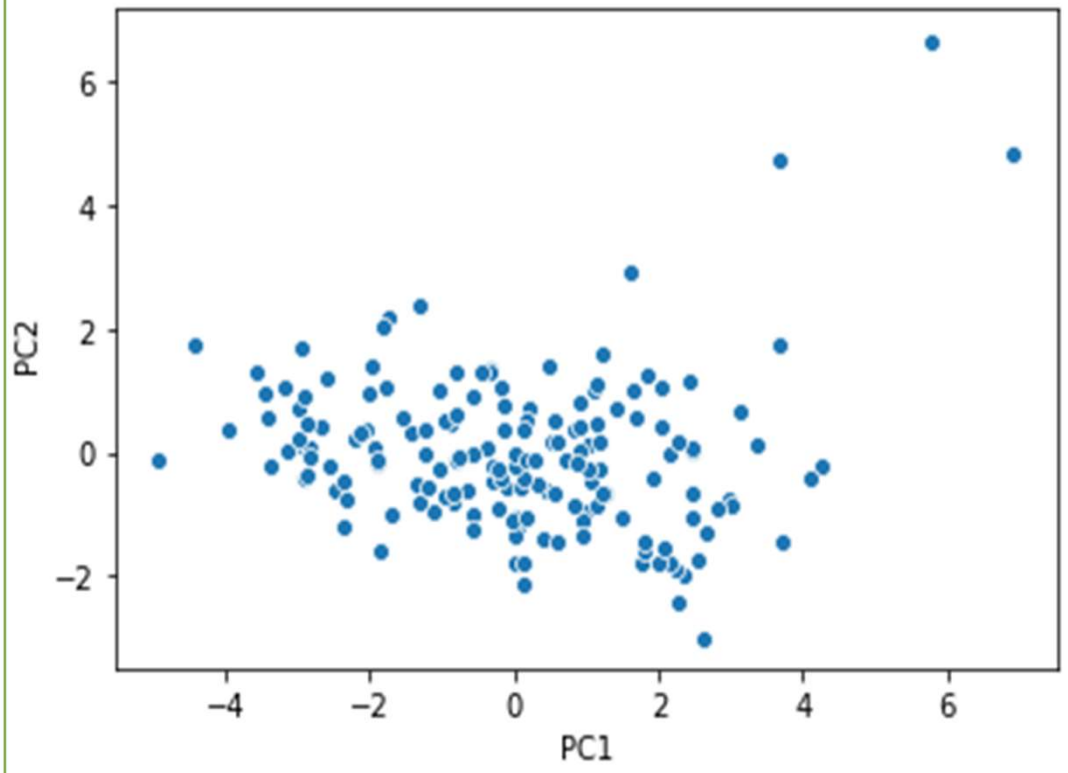
There is no multi collinearity.

The principal components are not correlated with each other.

The principal components are linear combination of the original variables. Thus, we also preserve most of the information, as we have not dropped any columns.

Scatter Plot - PC1 vs. PC2

The first two components that explain most of the variance are plotted.



Hopkins statistics for this data is greater than 0.7. The data has good tendency to form clusters.

Clustering

In this section, we apply the clustering algorithms:

K-means clustering and Hierarchical clustering.

K-means Clustering

Clustering is an unsupervised machine learning technique, where we try to find patterns based on similarities in the data.

The dataset is divided into groups called clusters such that the data points in each cluster are as similar (or close) as possible to one another (intra-cluster homogeneity) and the data points in different groups are as dissimilar as possible from each other (inter-cluster heterogeneity).

K-means algorithm is an iterative algorithm that tries to partition the dataset into K distinct non-overlapping groups (clusters) where each data point belongs to only one cluster.

Steps in K-means Clustering

- **Initialization** (randomly choose k points as centroids)
- **Assignment** (assign the different points to the nearest centroid) and
- **Optimization** (compute the centroid as an average of the points in the cluster). The assignment and optimization steps are repeated till the solution converges (i.e. the centroid do not change further)

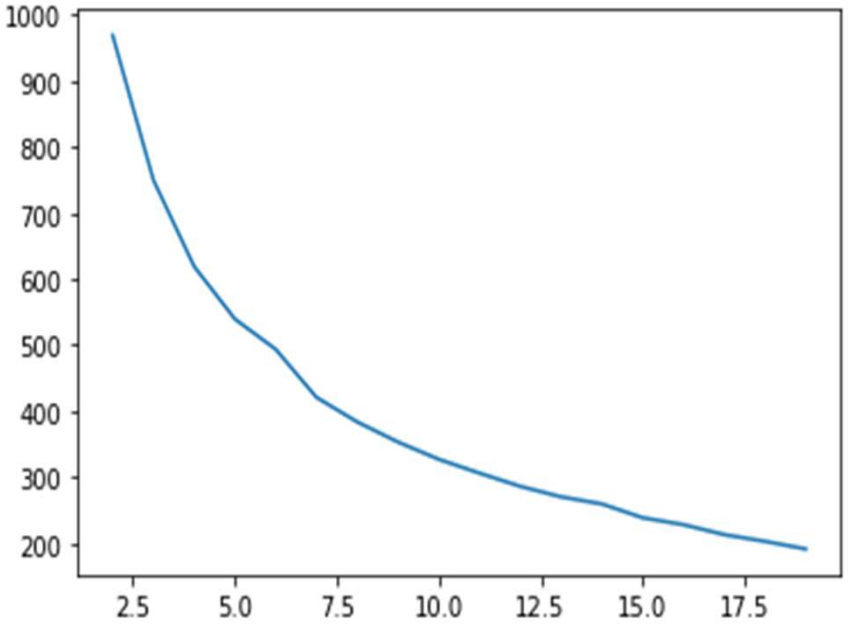
However, the first challenge is to decide how many clusters to be formed i.e. what is the value of k?

Finding the optimal value for K

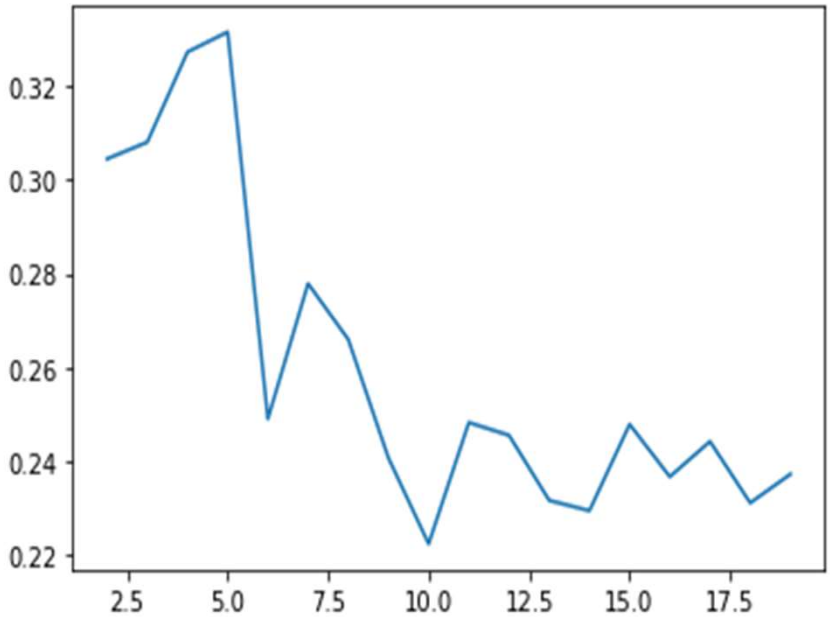
- **Elbow Curve:** For different values of k (2, 3, 4 …. 10 and so on) we apply the algorithm and then compute the sum of the squared distances (SSD) of the points to their closest cluster center. We plot these SSD against the number of clusters, which gives us the elbow curve. We select the point at which the marginal decrease (or the rate of drop) in the SSD value starts to diminish, which means that adding a new cluster is not adding any value the model.
- **Silhouette score:** The silhouette of a data point is a measure of how similar it is to data points within its cluster and how dissimilar it is matched to the data points in other clusters. Let q be the mean of the intra-cluster distance of a data point to all the points in its own cluster and p be the mean of the inter-cluster distance of the data point to the points in the nearest cluster that the data point is not part of. The silhouette score is computed as **(p - q) / max(p, q)**. The value of this score lies between 1 and -1. A score closer to 1 indicates that the point is very similar to other data points in its own cluster, whereas a score closer to -1 indicates intra-cluster dissimilarity.

Finding the optimal value for K

Elbow Curve

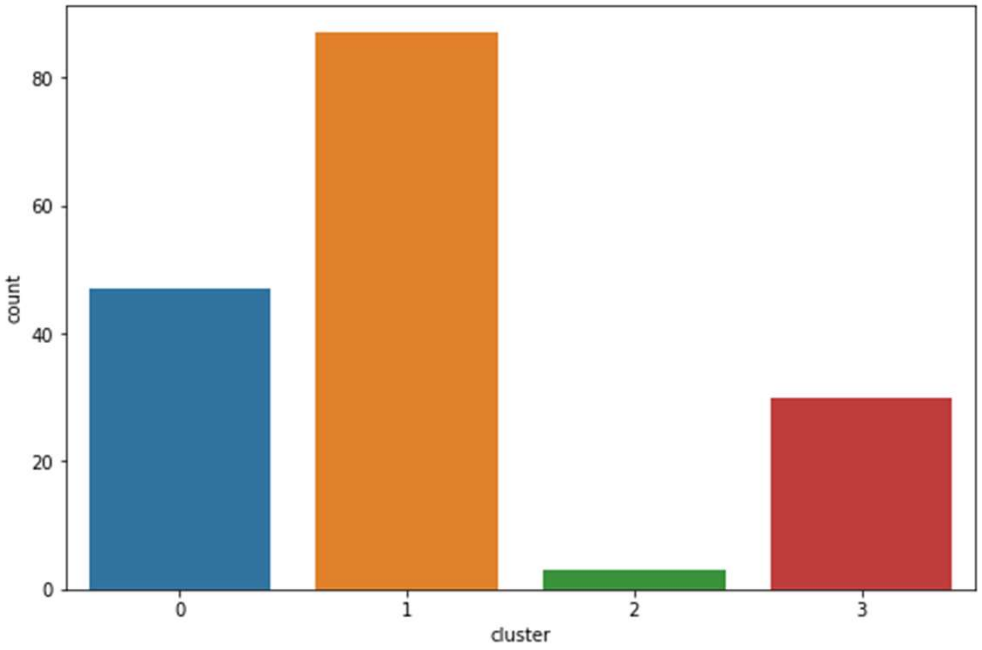


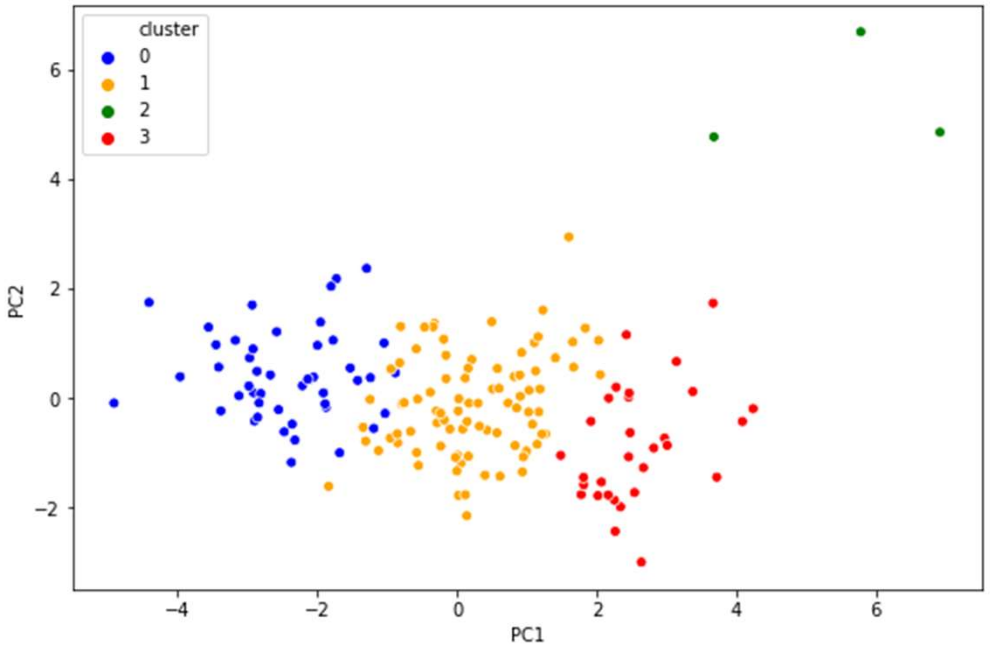
Silhouette Analysis



Based on the above analysis we can consider the optimum number of clusters to be 4 or 5. But when we do with 5 clusters, there is only one country getting into a fifth cluster. So we will choose to create only 4 clusters.

Creating the Clusters



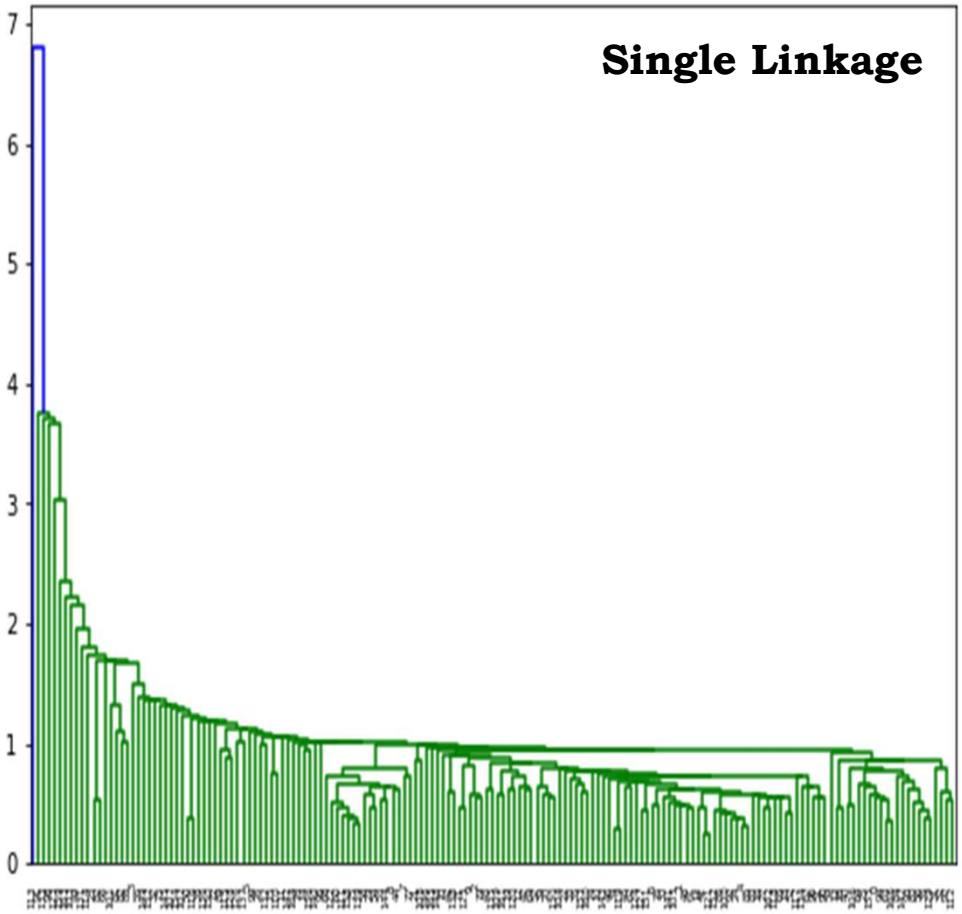


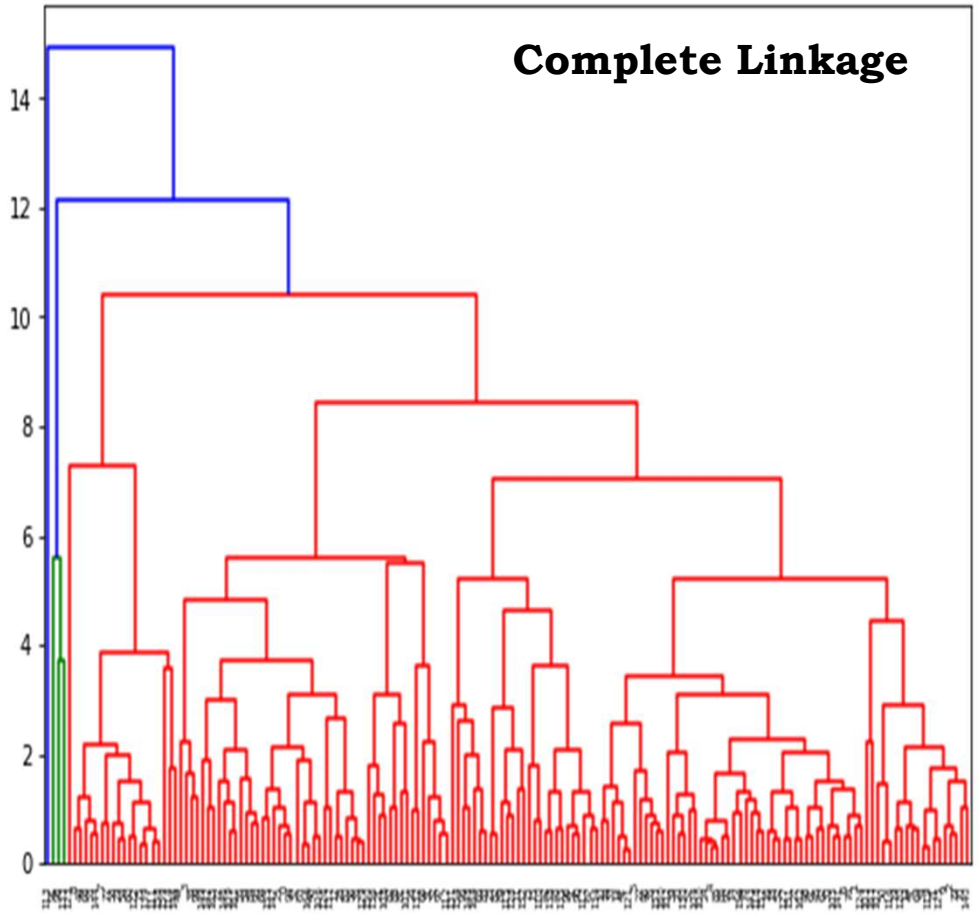
- **Cluster 0:** Blue Dots - Under Developed Countries
- **Cluster 1:** Orange Dots - Developing Countries (China, Russia, Malaysia, India and so on)
- **Cluster 2:** Green Dots - Some of the outliers (Luxembourg, Malta, Singapore)
- **Cluster 3:** Red Dots - Developed Countries (Norway, Switzerland, USA, Germany, Ireland, Australia etc.)

Hierarchical Clustering

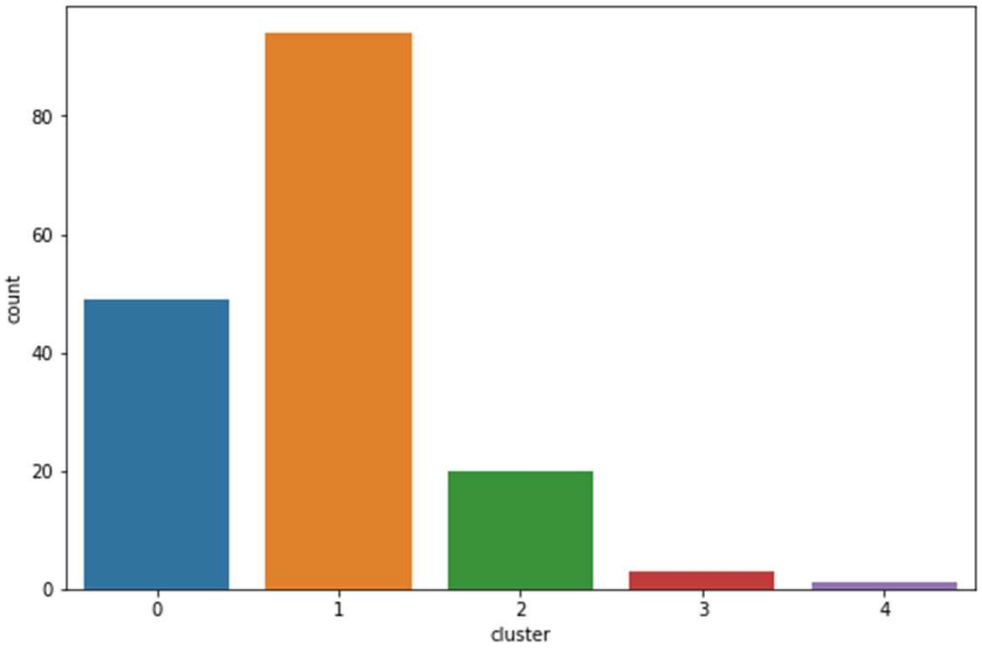
- In Hierarchical Clustering, we use the agglomerative approach to start with as many clusters as there are data points and then start merging the clusters that are closest to each other. There are three methods to compute the distance between two clusters.
- **Single Linkage:** The distance is computed as the shortest distance between points in the two clusters. If we represent the points in cluster r as x_r and the points in cluster s as x_s, then in single linkage we compute the distance between the clusters as $\min(D(x_{ri}, x_{sj}))$.
- **Complete Linkage:** The distance is computed as the longest distance between points in the two clusters. If we represent the points in cluster r as x_r and the points in cluster s as x_s, then in single linkage we compute the distance between the clusters as $\max(D(x_{ri}, x_{sj}))$
- Dendrograms produced using single linkage may not be structured properly. So to get a proper tree-like structure we use complete linkage or average linkage.

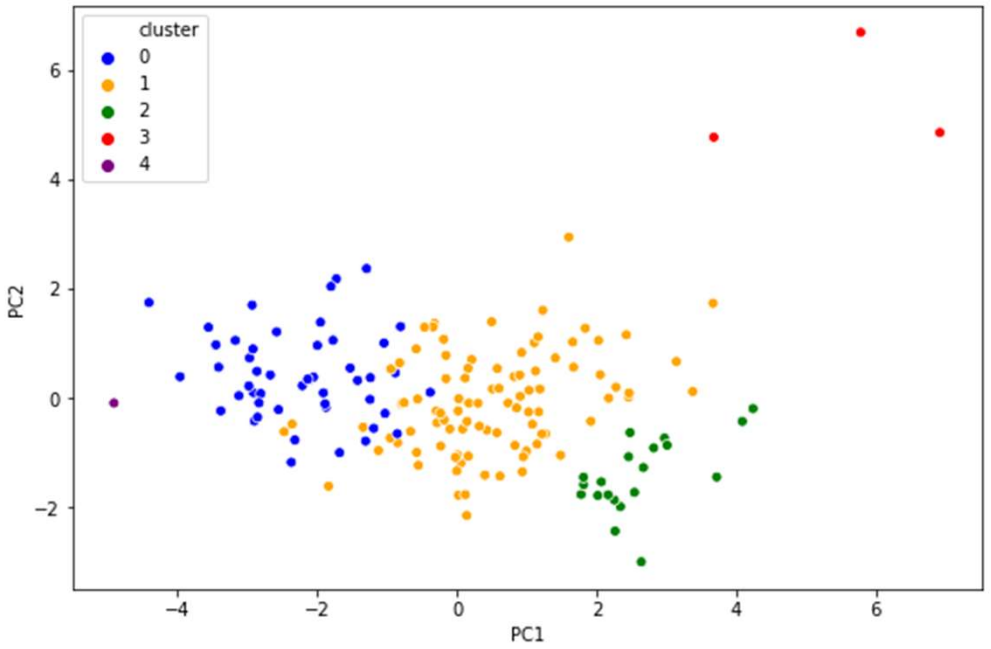
Dendrograms





Creating the Clusters



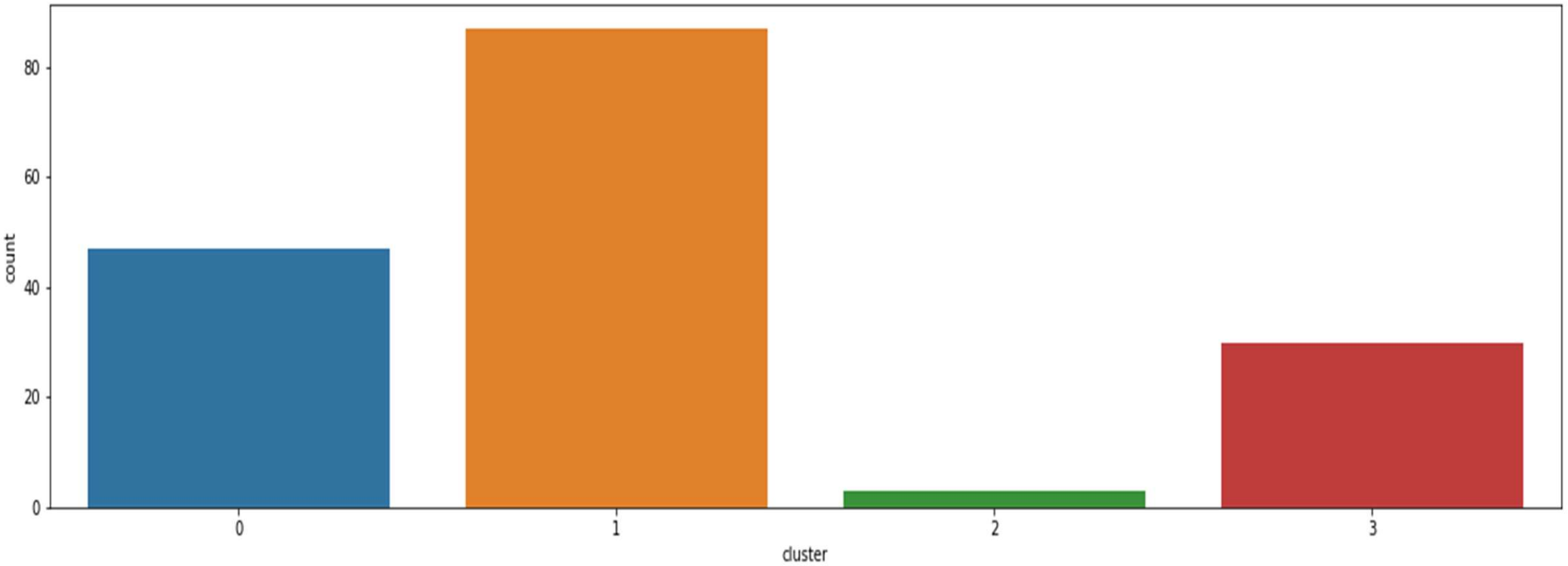


We got 4 clusters using hierarchical clustering which varies from the clustering given by K-means algorithm. But it produces better and more intuitive results compared to the k-means algorithm. However, for further analysis we will go with the clustering given by K-means.

Cluster Analysis and Recommendations

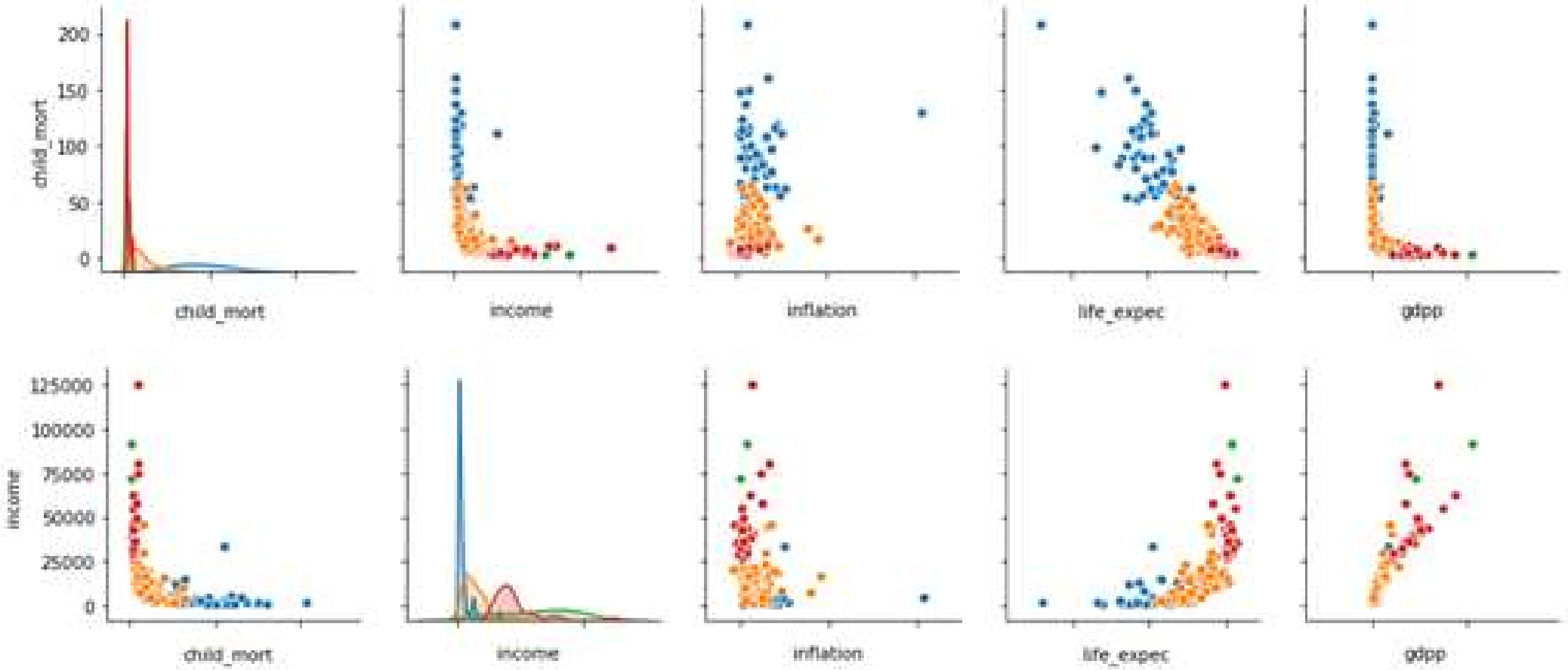
In this final and concluding section, we present a summary of our analysis and our recommendations on which are the countries that are in dire need for funding.

Clusters

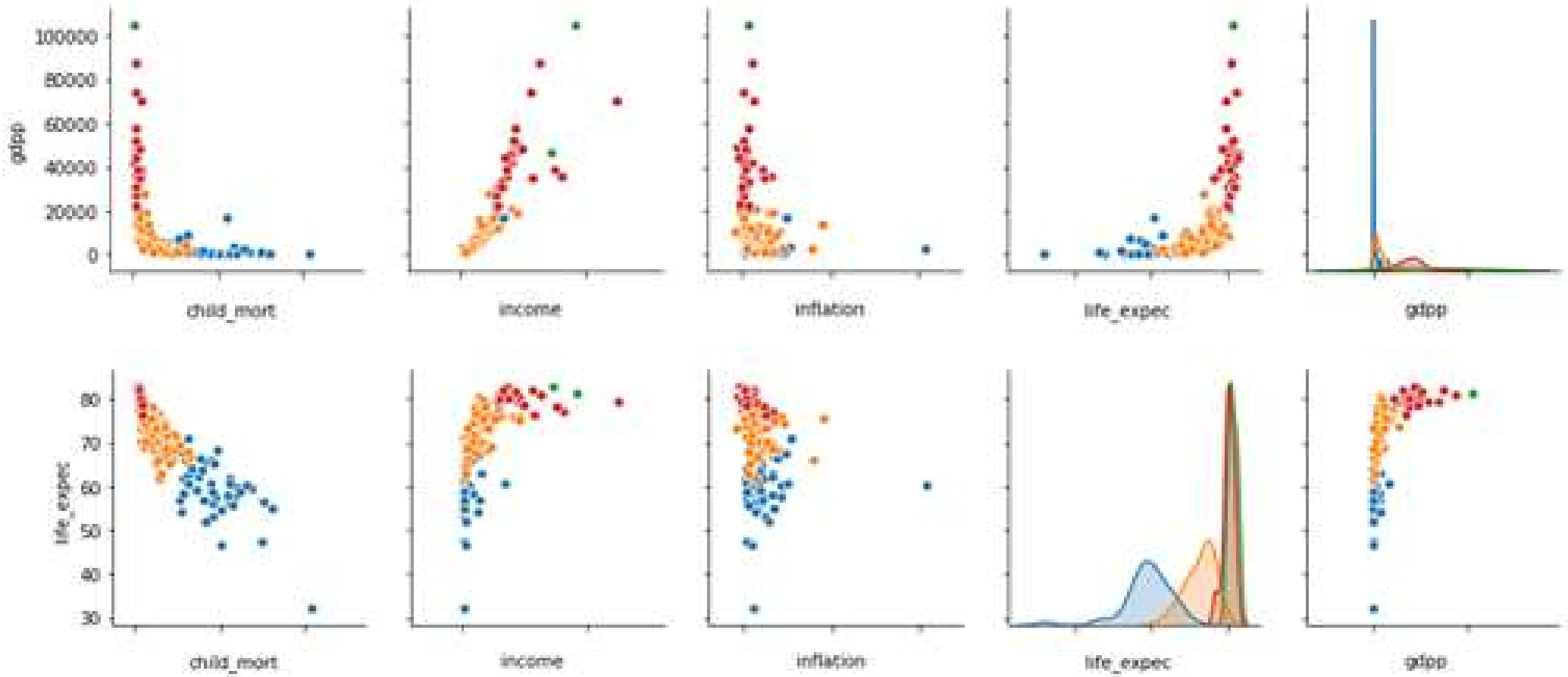


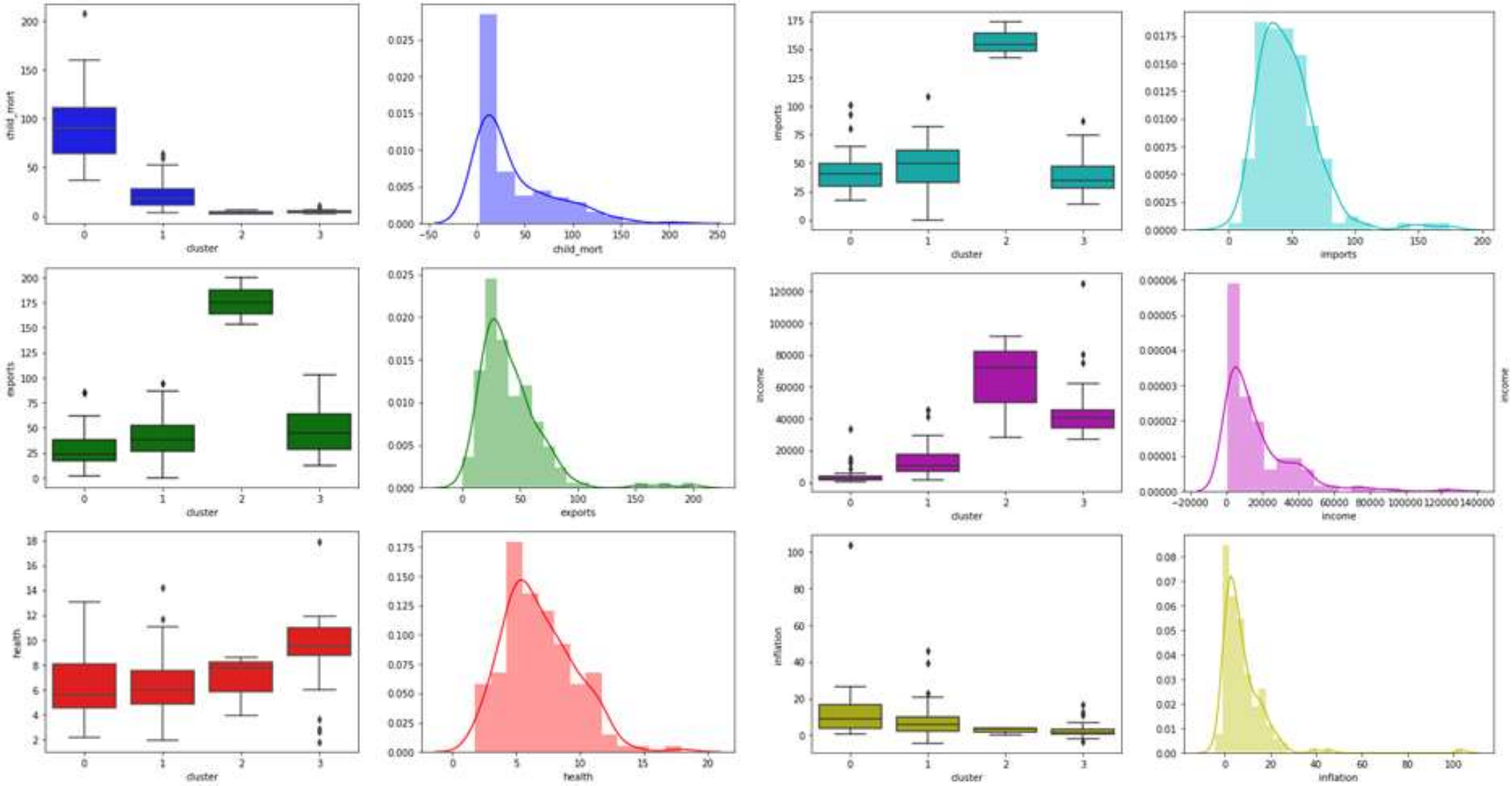
We created the clusters using the principal components. Now it is time to analyze the clusters thus formed using the original variables to identify the countries which you finally want to select.

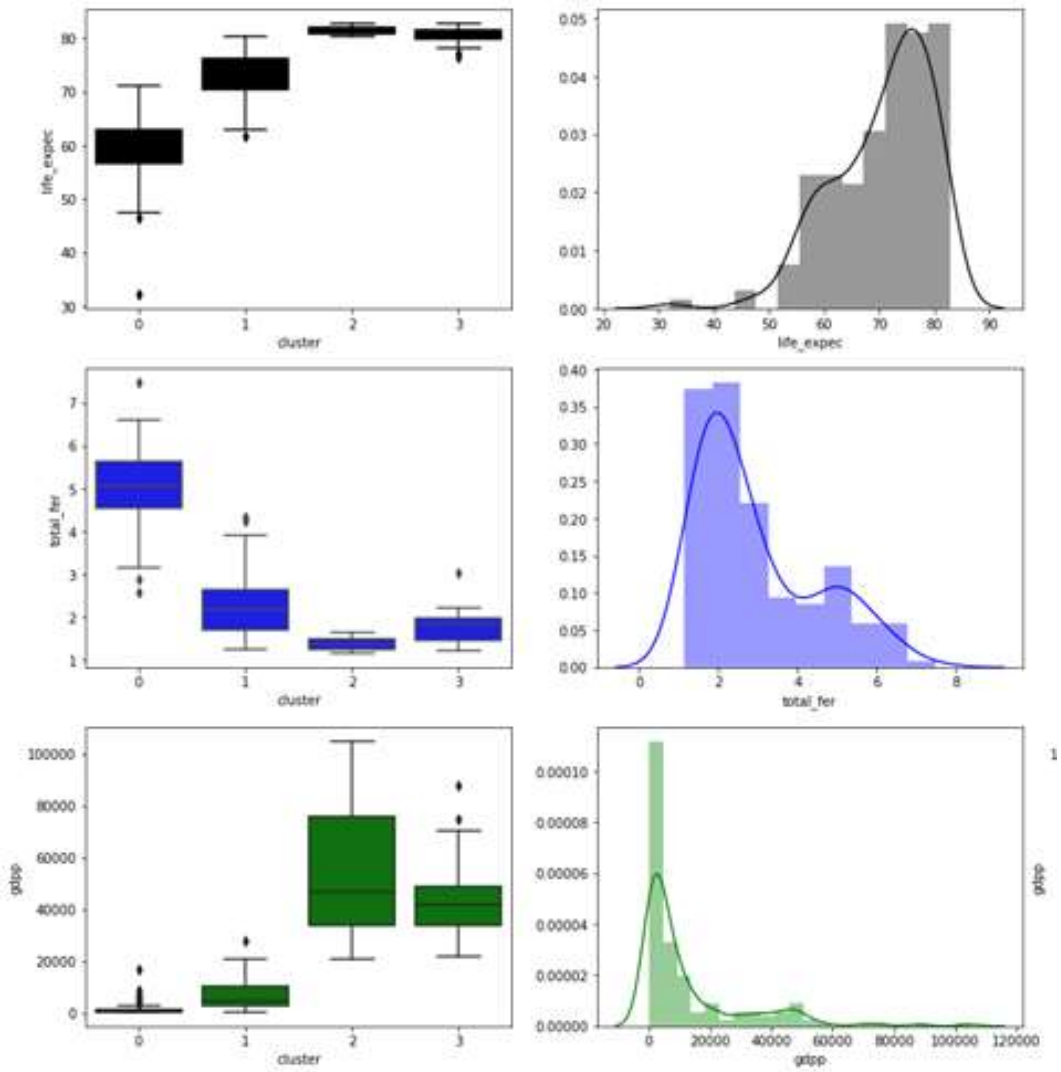
Visualizing the Clusters



Visualizing the Clusters




child_mort
cluster
child_mort
imports
cluster
imports
exports
cluster
exports
income
cluster
income
health
cluster
health
inflation
cluster
inflation



The box plots and distribution plots show how the data is distributed across the different clusters. The conclusions are presented in the next slide.

Cluster 0 - Under Developed Countries

From the above visualizations, we can see that, compared to other countries, the countries in cluster 0 have high child mortality rate, low life expectancy, low income and low gdpp, and high inflation rate. This is definitely a sign that these countries deserve more support and are in dire need for aid. These are under-developed countries.

Since most of the parameters are expressed in terms of the gdpp of the country, we will consider gdpp to rank the countries in cluster 0 to come up with the list of most deserving country. As we have seen earlier, the income is also strongly correlated with gdpp.

Recommendation

The countries which are having the dire need for funding aid are:

- Burundi
- Liberia
- Congo, Dem. Rep.
- Niger
- Sierra Leone
- Madagascar
- Mozambique
- Central African Republic
- Malawi
- Eritrea

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	cluster
26	Burundi	93.6	8.92	11.60	39.2	764	12.300	57.7	6.26	231	0
88	Liberia	89.3	19.10	11.80	92.6	700	5.470	60.8	5.02	327	0
37	Congo, Dem. Rep.	116.0	41.10	7.91	49.6	609	20.800	57.5	6.54	334	0
112	Niger	123.0	22.20	5.16	49.1	814	2.550	58.8	7.49	348	0
132	Sierra Leone	160.0	16.80	13.10	34.5	1220	17.200	55.0	5.20	399	0
93	Madagascar	62.2	25.00	3.77	43.0	1390	8.790	60.8	4.60	413	0
106	Mozambique	101.0	31.50	5.21	46.2	918	7.640	54.5	5.56	419	0
31	Central African Republic	149.0	11.80	3.98	26.5	888	2.010	47.5	5.21	446	0
94	Malawi	90.5	22.80	6.59	34.9	1030	12.100	53.1	5.31	459	0
50	Eritrea	55.2	4.79	2.66	23.3	1420	11.600	61.7	4.61	482	0

Thank You

Submitted by: Janarthanan B