॥ त्वं ज्ञानमयो विज्ञानमयोऽसि ॥

PRML Minor Project

Report

Project :- 05

Team Members

1.Abhishek Arya (B21AI002)

2.Lavish Gupta (B21AI020)

3.Pankaj Kumar (B21AI024)

Dataset Used:- Country_Data

Dataset Overview

The Country_Data dataset consists of socio-economic and health data for 167 countries. The data covers indicators such as GDP per capita, child mortality rate, and percentage of the population below the poverty line, among others.

Description:-

1. This report provides a data-driven analysis of which countries are in the direst need of aid based on socio-economic and health factors.
2. The report uses the Country_Data dataset, which covers indicators such as GDP per capita, child mortality rate, and access to electricity for 167 countries.
3. The report performs data cleaning and uses principal component analysis (PCA) to identify the most significant factors that contribute to a country's development.
4. The report uses k-means clustering to group the countries into four categories: least developed, less developed, moderately developed, and developed.
5. The report concludes that by targeting aid to these countries, HELP International can make a meaningful impact on global development and improve the lives of millions of people.

Objective:

The objective of this project is to help the CEO of HELP International to make informed decisions about how to allocate $10 million in aid by categorizing countries based on socio-economic and health factors that determine their overall development.

1.Reading the dataset(Data Preprocessing)

1. Imports necessary libraries such as pandas and numpy.
2. Reads the CSV file named "Country-data (1).csv" into a pandas dataframe named 'dataset'.
3. Prints the first 20 rows of the dataframe using the `head()` function.
4. Prints a summary of the dataset using the `describe()` function.
5. Converts the 'exports', 'imports', and 'health' columns into values of contribution in GDB.
6. Prints the shape of the dataframe using the `shape` attribute.
7. Checks for missing values in the dataset using the `isnull()` function.
8. Drops the 'country' column from the dataset using the `drop()` function

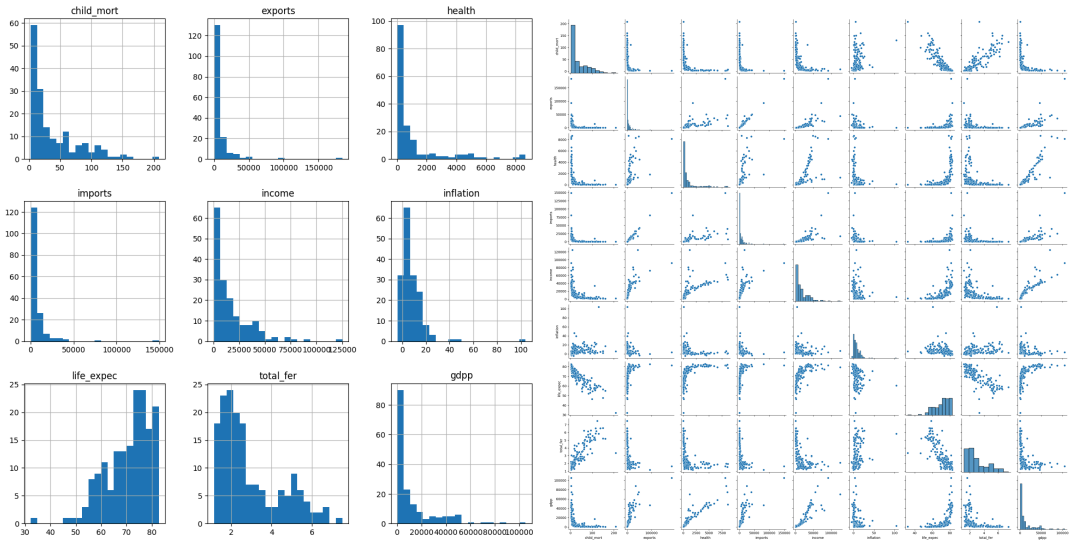
2.Dataset Visualisation

Visualization techniques to better understand the dataset and apply a preprocessing step of data scaling to prepare the data for further analysis.

a. Visualizes the dataset using various plotting techniques:-

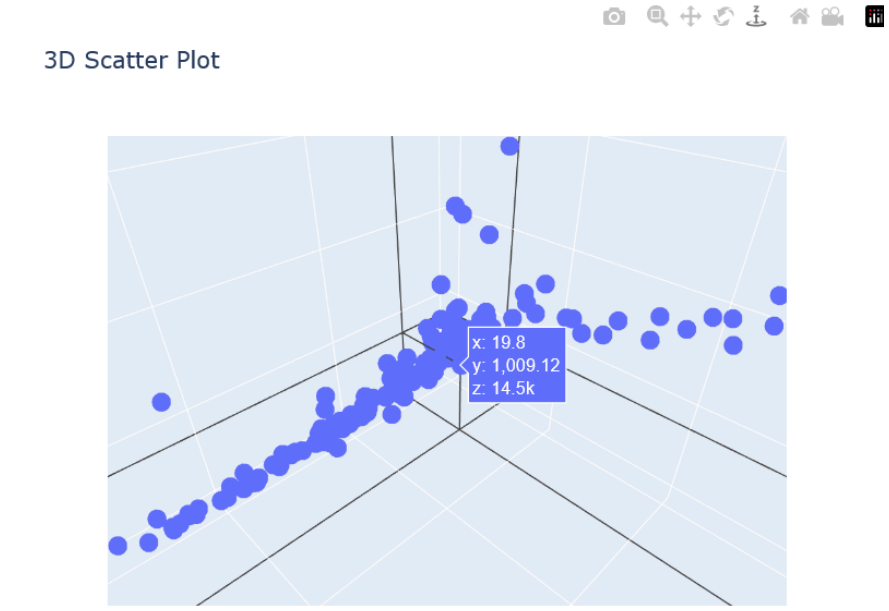
1. Histogram and Seaborn

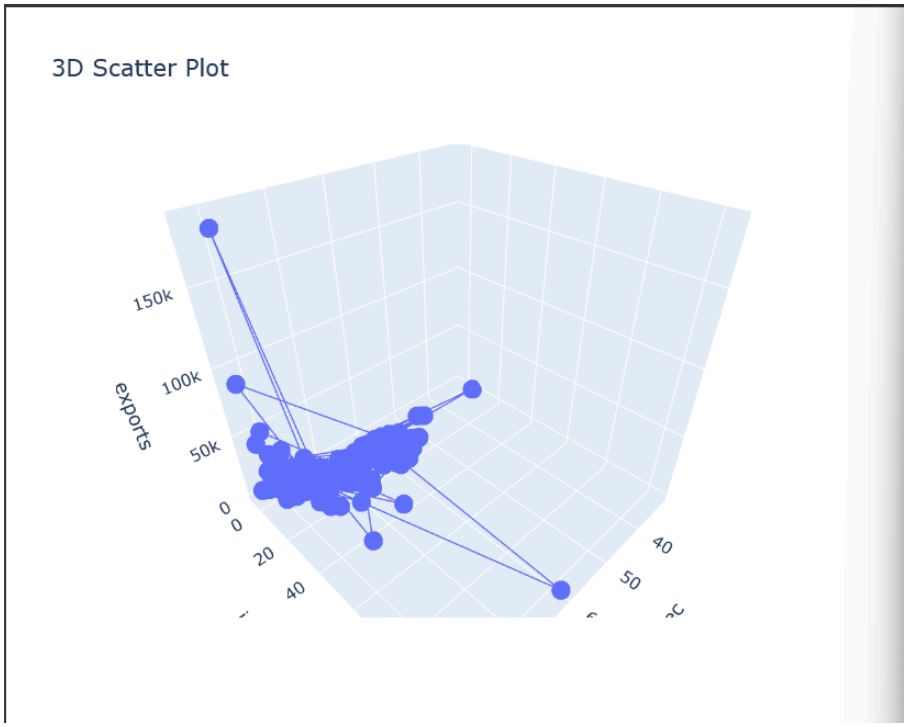
Plots a histogram of the dataset and a scatterplot matrix using **Matplotlib** and Seaborn libraries. The scatterplot matrix shows the scatterplots between each pair of variables in the dataset, while the histogram shows the distribution of each variable.



2. 3D scatterplots

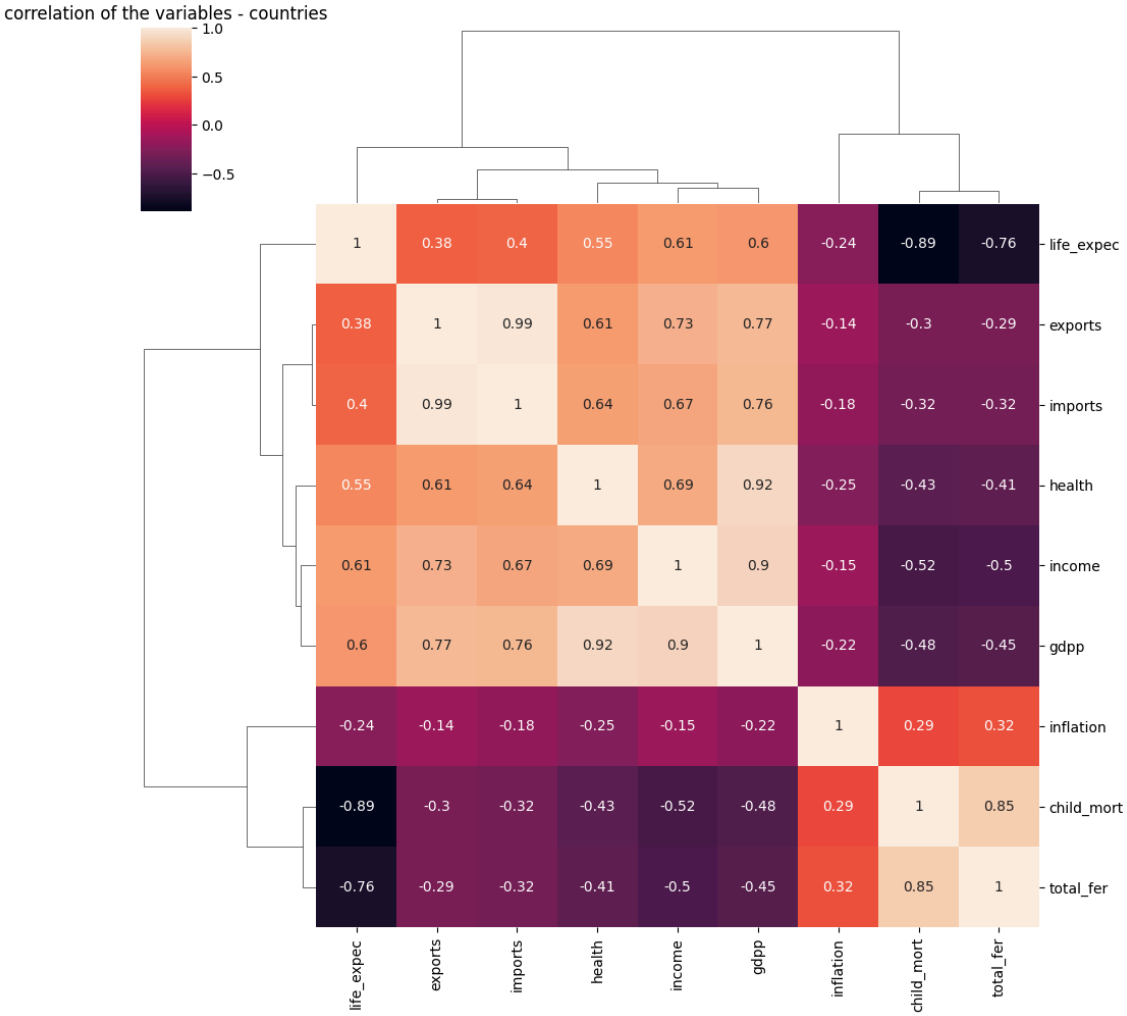
The first scatterplot shows the relationship between child mortality, health, and income, while the second scatterplot shows the relationship between life expectancy, inflation, and exports.


3D Scatter Plot
x: 19.8
y: 1,009.12
z: 14.5k


3D Scatter Plot
150k
100k
50k
0
exports
0
20
40
40
50

3. Clustermap

Uses Seaborn library to create a clustermap of the **correlation matrix** of the variables in the dataset. This clustermap visualizes the strength and direction of the correlations between the variables.



->Using StandardScaler from Scikit-learn library, standardizes the data by subtracting the mean and dividing by the standard deviation of each variable.

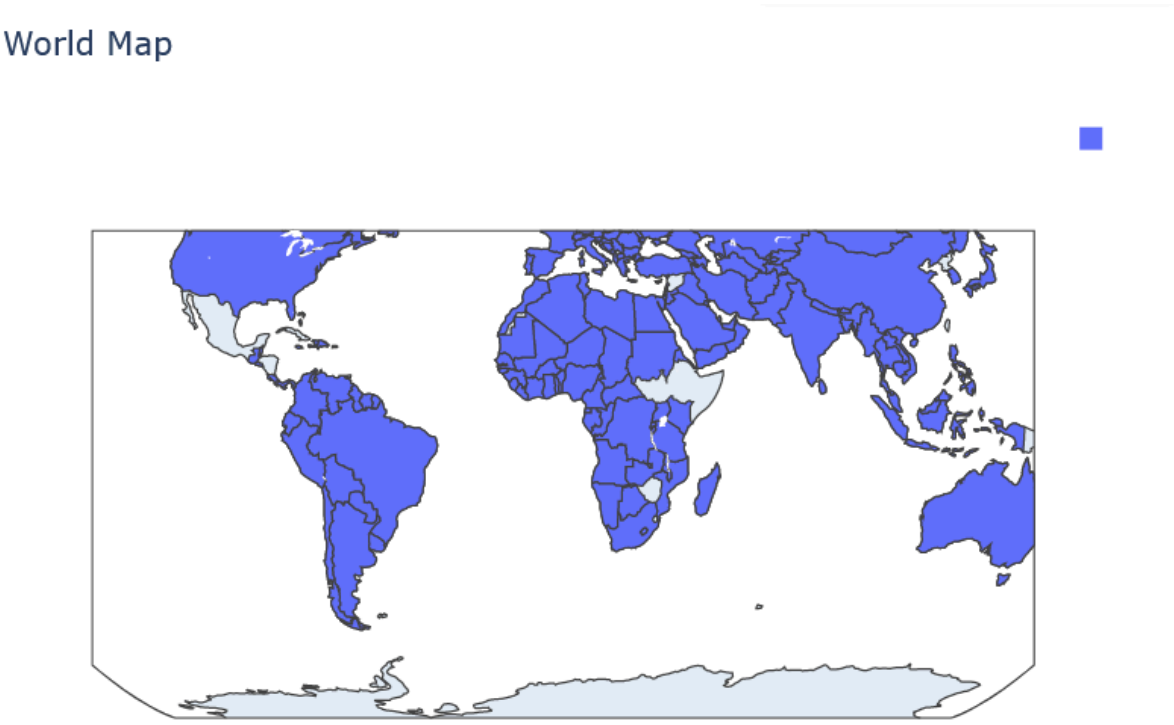
Note:-

1. The heatmap shows that some variables have high correlations with positive and negative responses, which need to be removed.
2. PCA will be used to overcome this multicollinearity, preserving valuable information and reducing dimensionality. This will take care of multicollinearity while also preserving valuable information and reducing dimensionality.

4.Choropleth map of the world.

The dataset1 contains a column named country, which is used to match the country names in the df DataFrame with their corresponding locations on the map.

The hover_name parameter specifies the text that appears when hovering over a country, and the resulting map is displayed using fig.show().



3.Dimensionality Reduction-PCA Components(PC1 vs PC2)

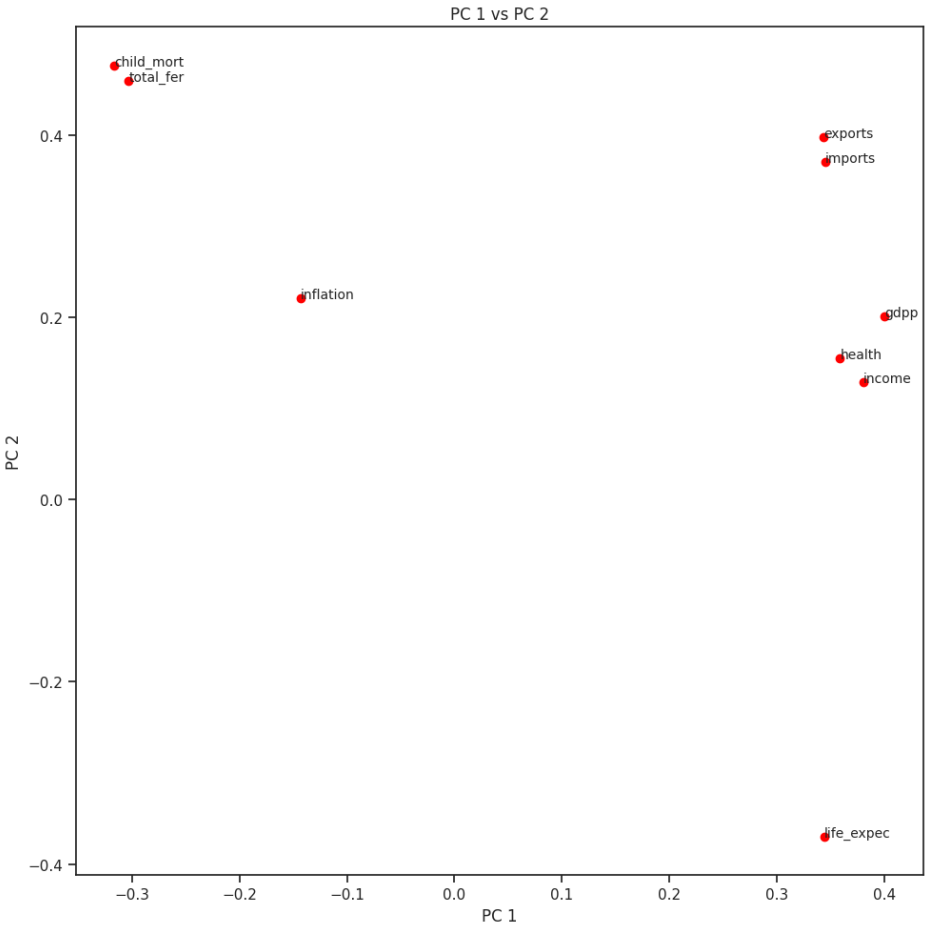
Principal Component Analysis (PCA) on a dataset containing information on several countries. Principal components use to visualize and analyze the data, and to identify patterns or relationships between the variables.

PCA is performed on the scaled data using the PCA function from sklearn.decomposition.

1. PCA components(PC1 vs PC2).

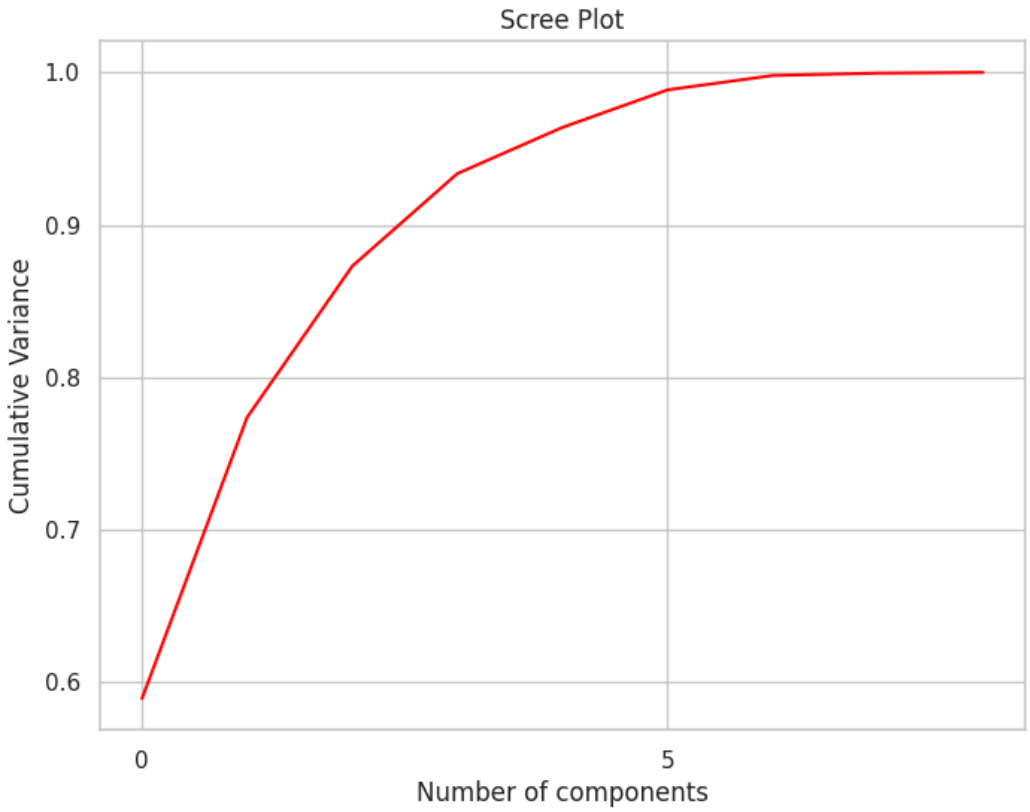
a. The data points of life expectancy is in the direction of principal component 1.
b. All the high data points are in the direction of principal component 2.

c. Labels each point with the name of the corresponding feature.



2.PCA- Scree plot.

a. Created a scree plot to show the explained variance of each principal component, and identifies the number of components needed to explain a certain percentage of the total variance.



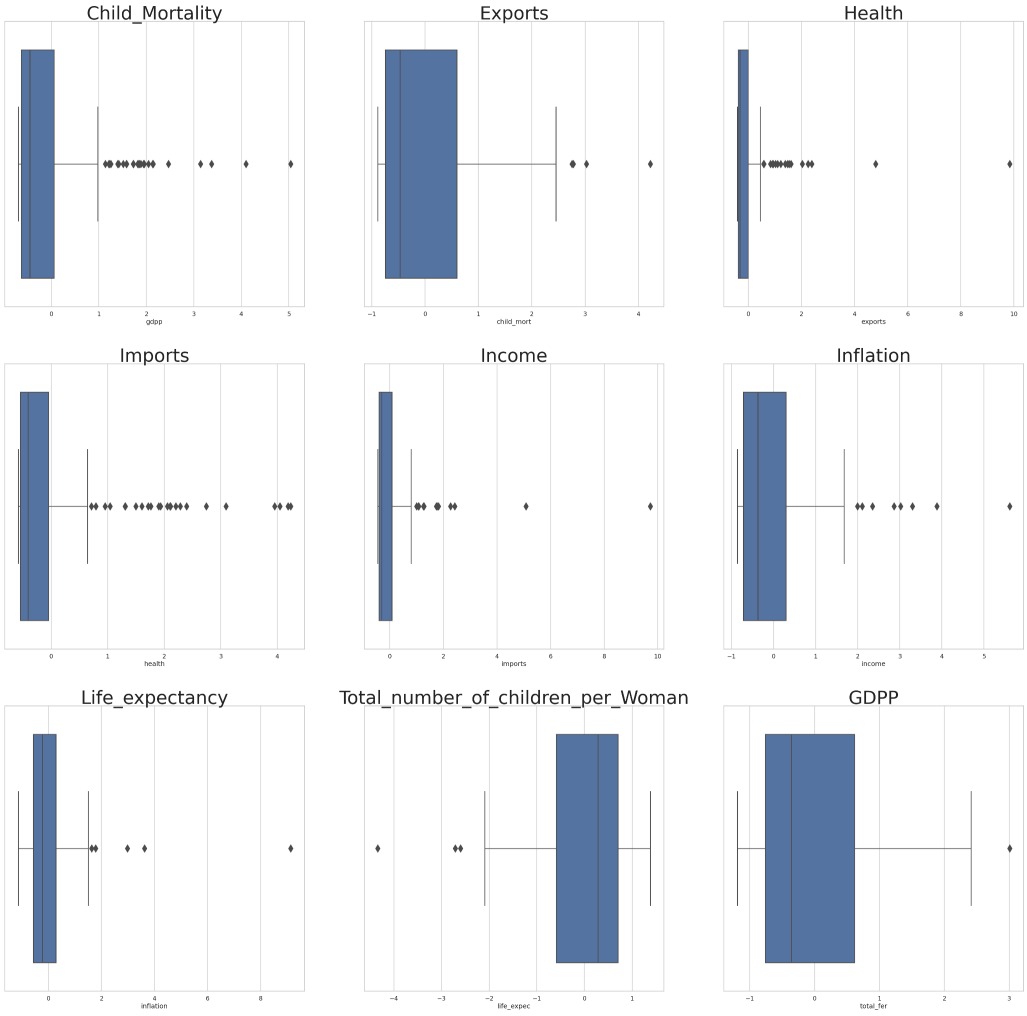
From graph we can see that:-

a. The number of component equal to 4 is having approx 95% of variance explained
b. The number of components equal to 5 is having approx 97% of variance explained.
c. So, the ideal number of components can be chosen is 5.

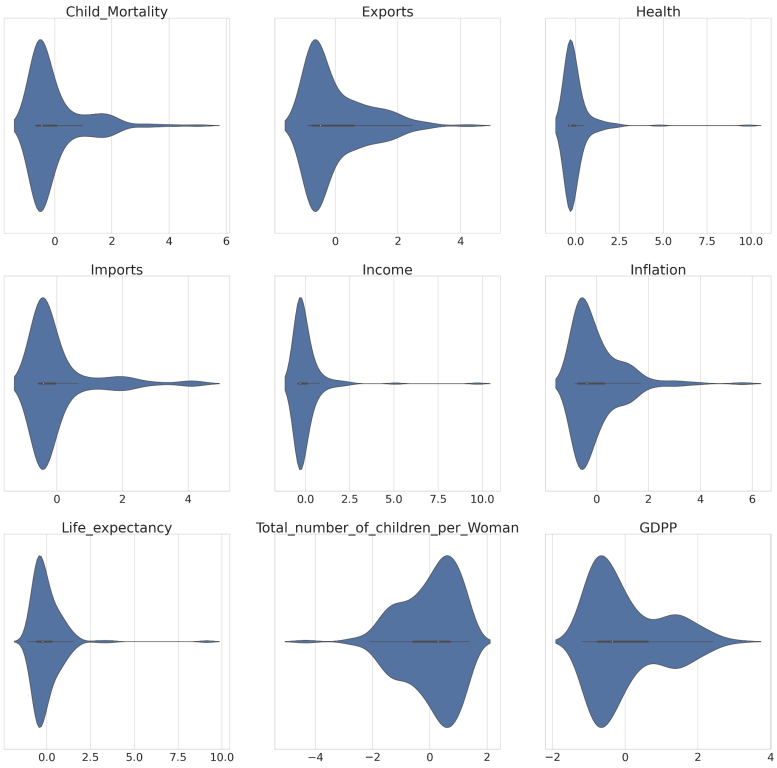
3.Visualization of Outliers.

Box plot and violin plots to visualize the spread and outliers of each variable in the original dataset, before and after scaling.

1.Boxplot.



2.Violin plot



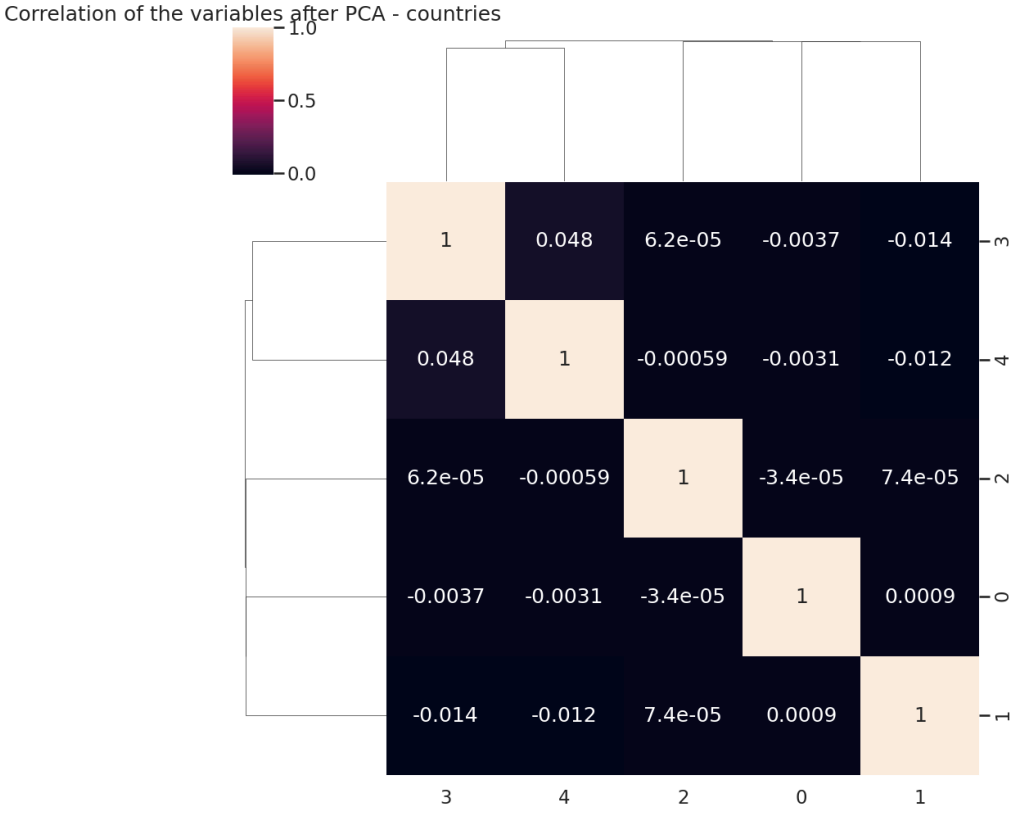
->Information we get form the graph:-

a. All the boxplots are having decent amounts of outliers.
b. All boxplots expect one ‘Total_number_of_childre_per_women’ is having outliers on the bottom of the boxplots which means there are some countries were no of children per women is very less compared to the other countries.
c. The inflation boxplot is having very thin size of quartiles compared to other countries.

4.Correlation after PCA

Performs Incremental PCA on the scaled data to reduce the number of dimensions to 5, and creates a correlation matrix of the reduced dataset.

The resulting correlation matrix is then plotted as a heatmap using the clustermap function from the seaborn library.



Note:-

Plots and datasets can be used to gain insights into the underlying patterns and relationships in the original data, and to identify potential outliers.

4.K-Means Analysis

K-means clustering algorithm to cluster the countries based on their similarity in terms of the socio-economic indicators.

Clusters_range=[2,3,4,5,6,7,8]

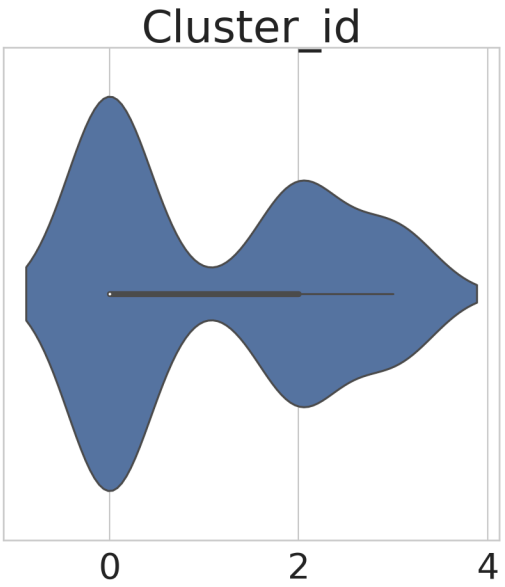
1. First determines the optimal number of clusters using the **silhouette score.**

```
Cluster=2 has silhouette score 0.480660467827559 17
Cluster=3 has silhouette score 0.45415128979715 08
Cluster=4 has silhouette score 0.46298146414826 11
Cluster=5 has silhouette score 0.46500328500357 124
Cluster=6 has silhouette score 0.43754410238617 975
Cluster=7 has silhouette score 0.35485432593701 65
Cluster=8 has silhouette score 0.36935653298132 49
```

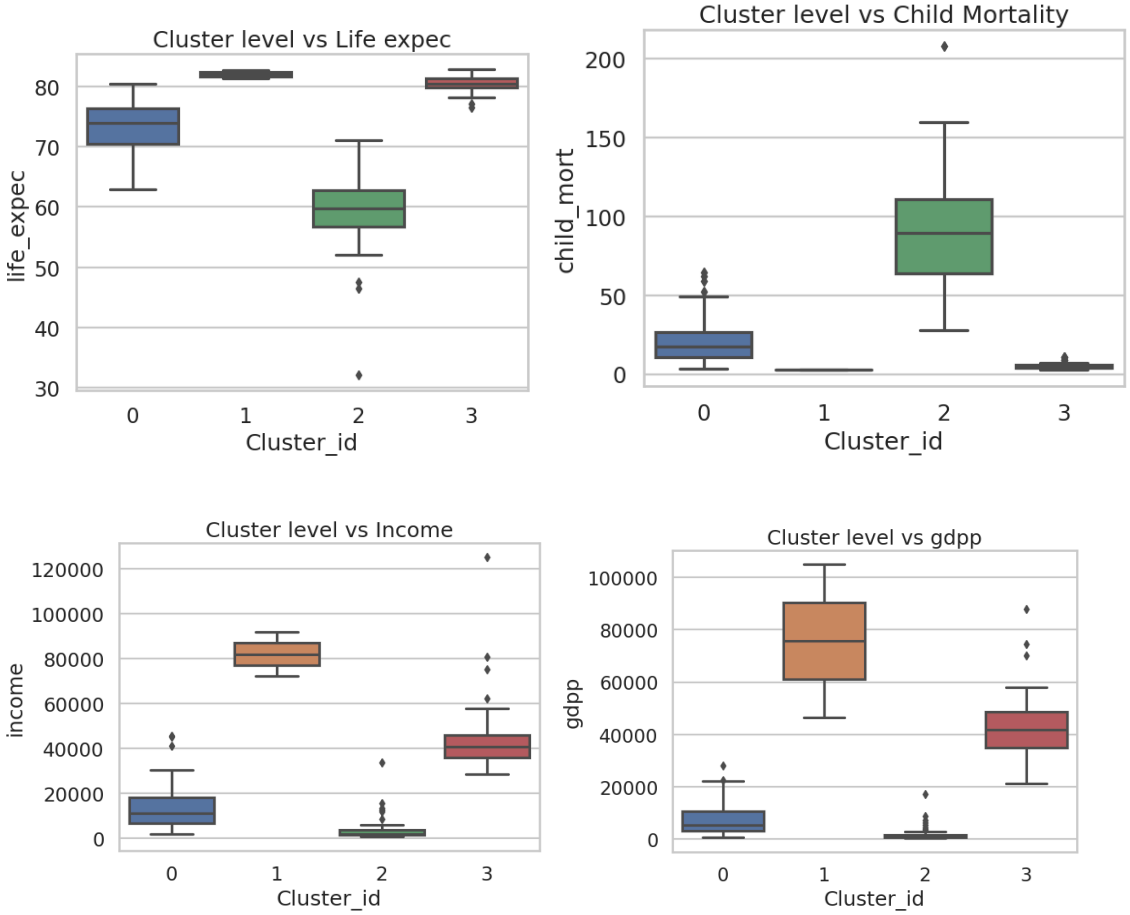
2. Seeing the silhouette score, we conclude that cluster =2 is having **highest score** hence k value should be 2 but,if we choose k value as '2', it will not suit business needs.
3. Therefore, we use k value as **'4'** since it is giving precise information also fulfills business needs.
4. Assigns a cluster label to each country.

1. Visualization of the original variables with clusters

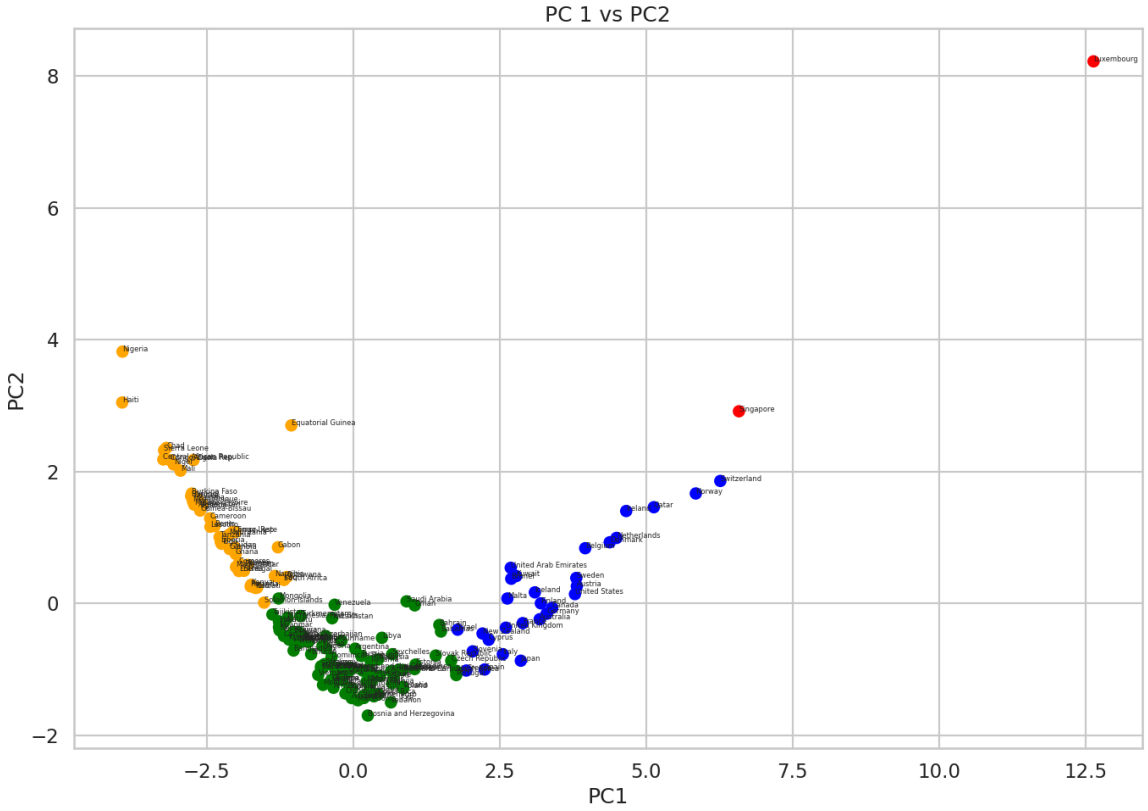
a. Plots the cluster distribution using a **violin plot.**



b. Using boxplot(relationship between the socio-economic indicators and the clusters)



c. Created a scatter plot to visualize the results of the PCA analysis and to show the clusters' separation.(PC component 1 and 2 in X-Y axis)



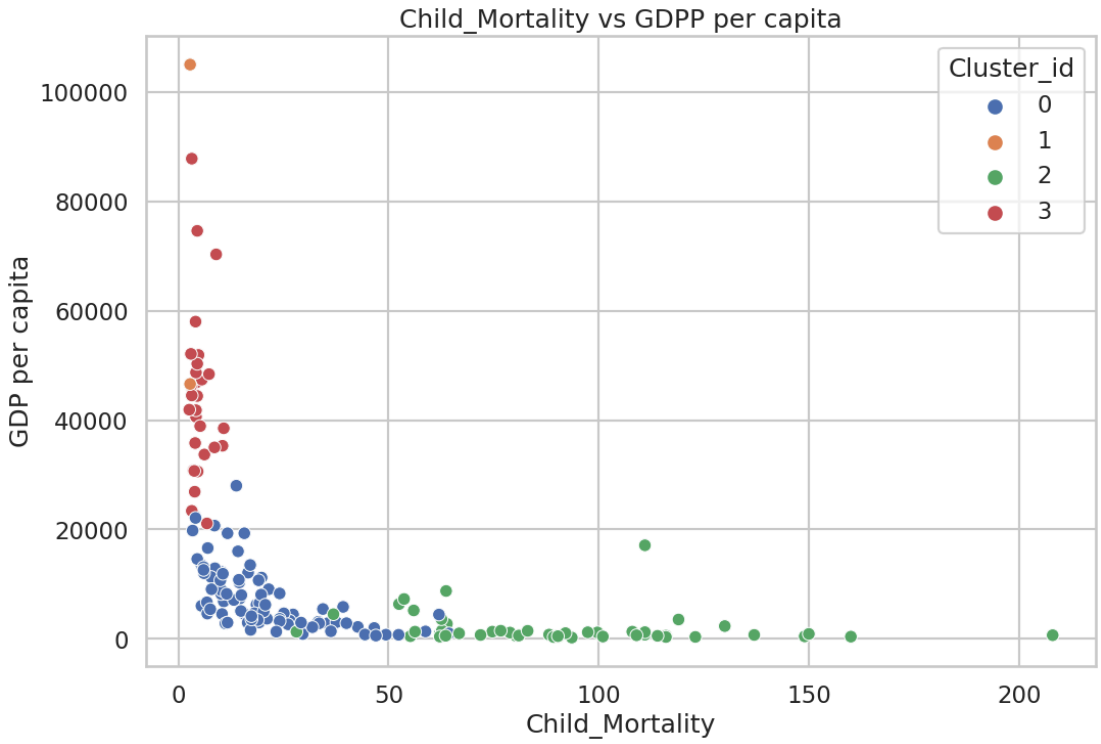
From the scatter plot, insight we get:-

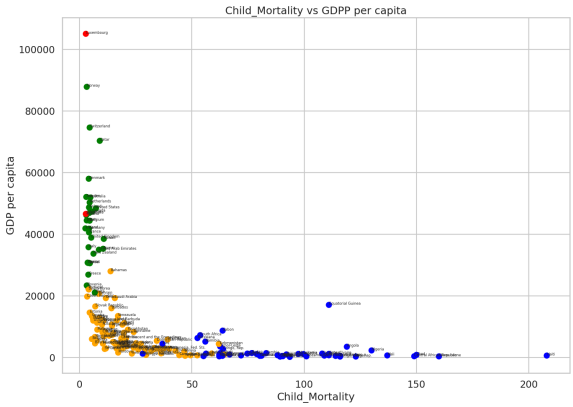
a. The PC1 is in the direction where the countries need of least help. Here, we choose **PC1** because it has **maximum percentage of variance** explained.
b. Countries like **Luxembourg and Singapore** are having high PC1 that means they are doing well,on the other hand **NIgeria, Hiti**,etc need urgent aid.
c. Therefore , **‘Orange’** are in direst need of aid than color in **‘Blue’.**

d. Child mortality vs GDPP per capita scatter plot.

1. Each point colored by its cluster label.

   The countries belonging to **cluster 0** are the most **underdeveloped** and require immediate assistance. **Cluster 1** consists of countries with **moderate development**, while **clusters 2** and **3** represent relatively **developed countries.**



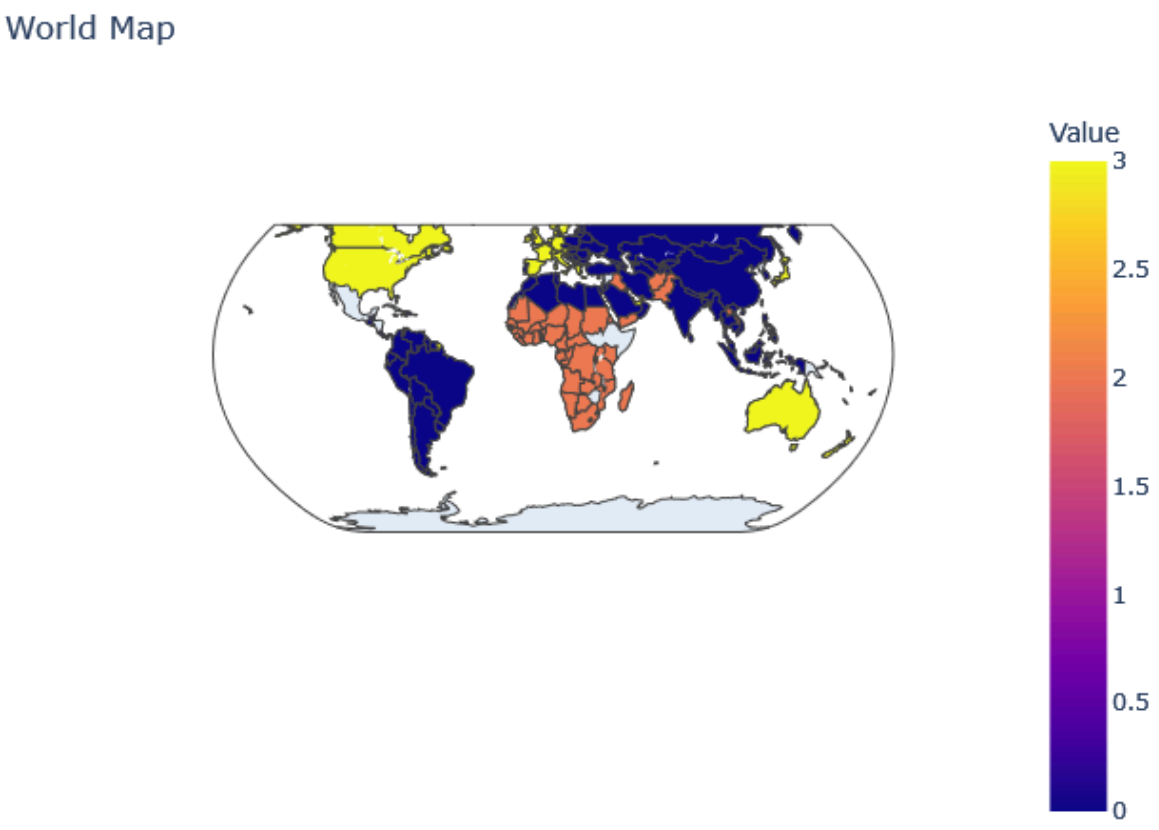


Insight we get from the graph:-

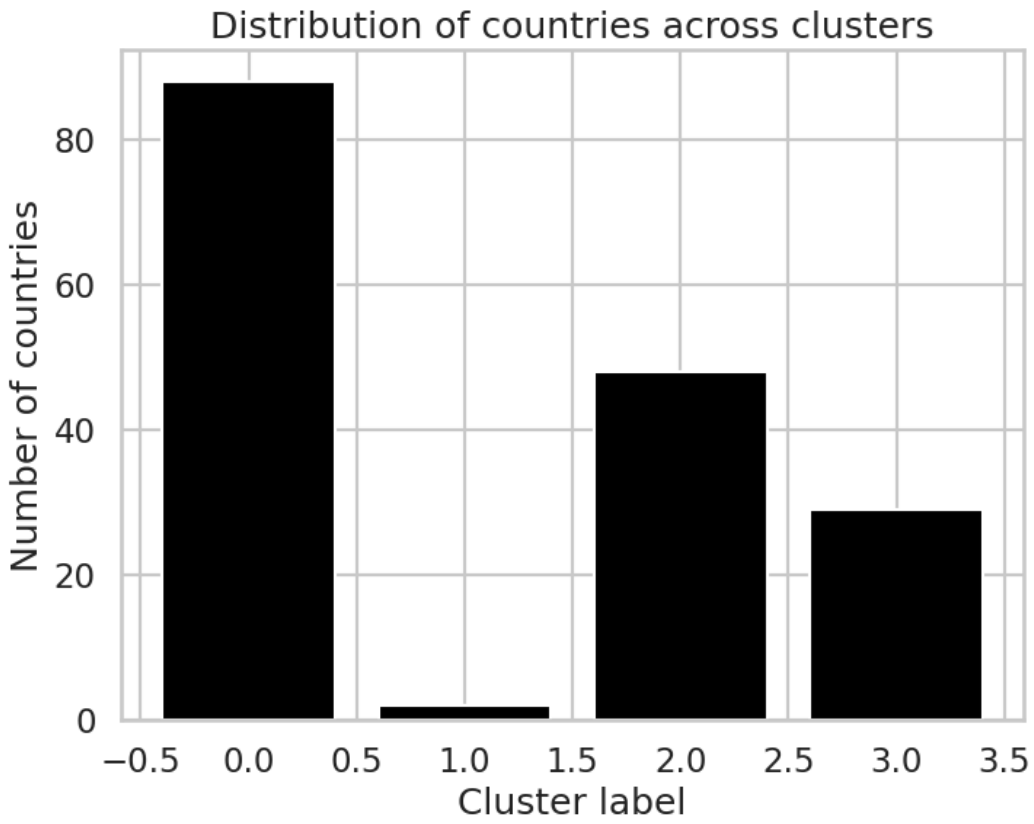
   a. Country like ‘Haiti’ is in need of aid

b. On the other hand ‘Luxembourg’ is having good gdpp and less child mortality.
c. These two are outliers.

E. Choropleth map using Plotly to visualize the distribution of the clusters across different countries.



F. Bar graph to show the distribution of number of country in each cluster.



Note:-

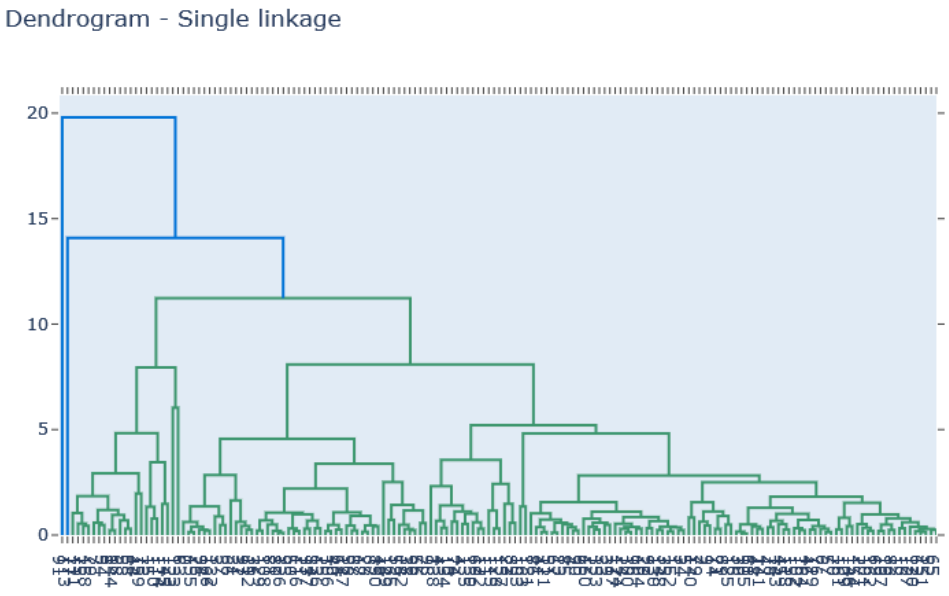
a. Clusters with label=0,1,2 and 3 corresponding to them there are 88,2,48 and 29 countries
b. Overall, These plots show how clustering analysis can be used to identify and categorize underdeveloped countries based on various socio-economic factors, which can help CEOs and organizations(NGO's) prioritize their aid and development efforts.
c. There are a total of 48 countries from the dataset that need urgent aid as they are having the lowest income, high child mortality and low gdp per capita.
d. 2 countries with good socio-economic and health factors.

5.Hierarchical clustering

Hierarchical clustering technique groups the countries with similar features together.

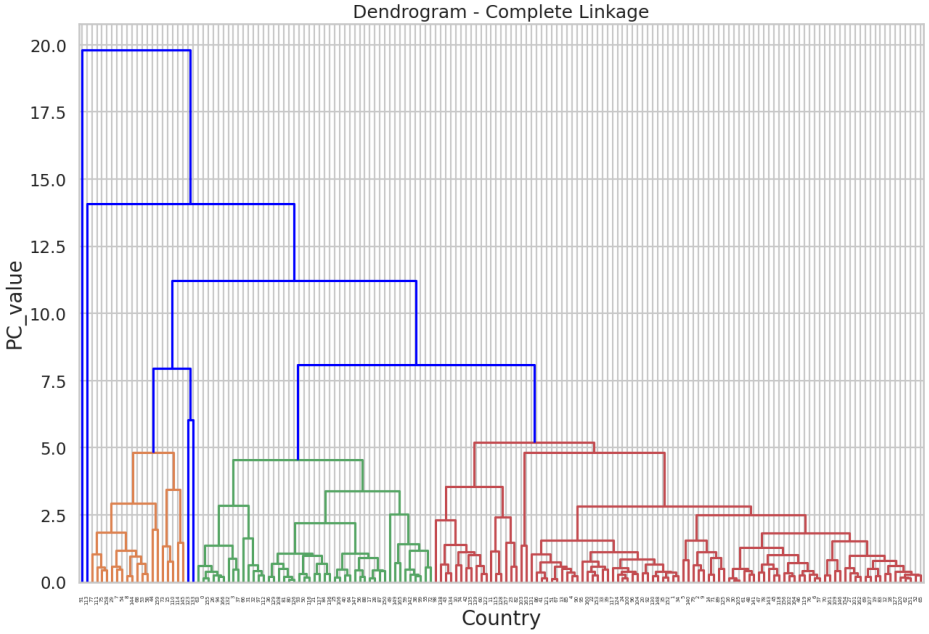
1.Single Linkage

This is another method to find out low development countries.



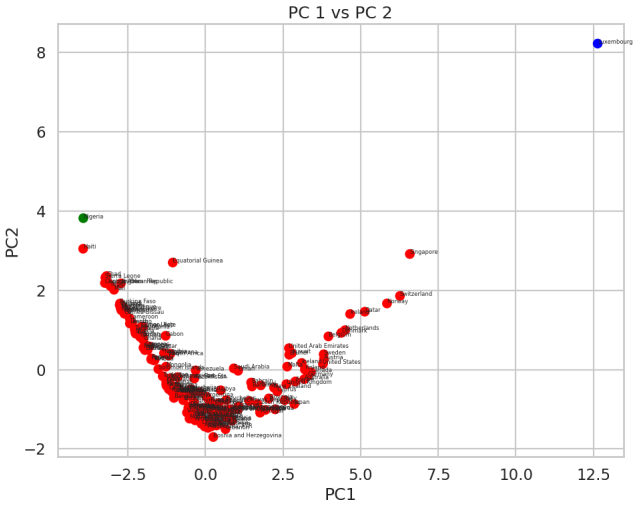
a. As we can see from the graph of linkage dendrogram, it is not quite visible and doesn't suits properly with dataset because we can cut the tree in a threshold value, we will use complete linkage dendrogram for hierarchical clustering.

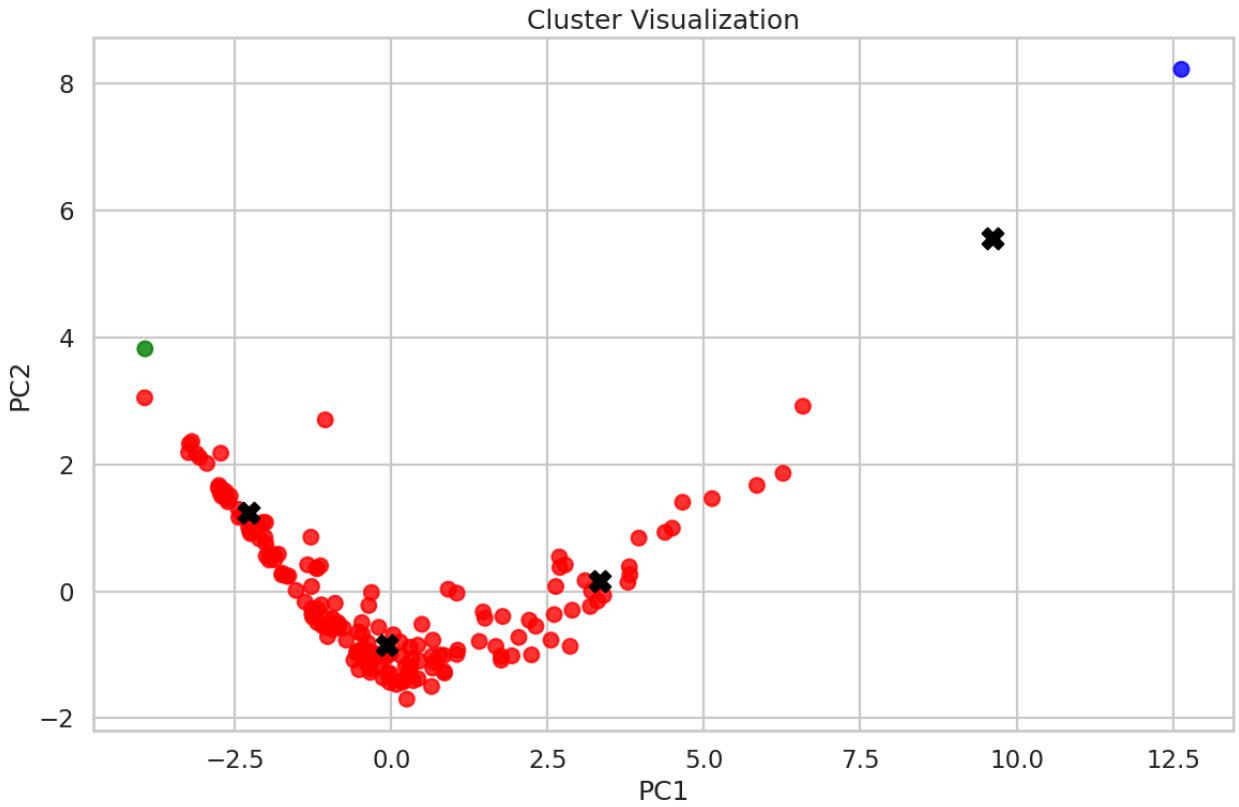
2.Complete Linkage



a. From the graph, we conclude that we will cut at 3 branches which will give us 3 clusters.
b. Therefore, the countries have been grouped into 3 clusters based on their socio-economic indicators

3.Visualization of hierarchical clustering with first two principal components.



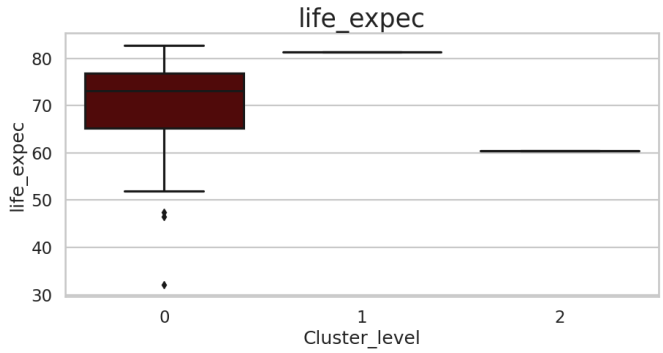


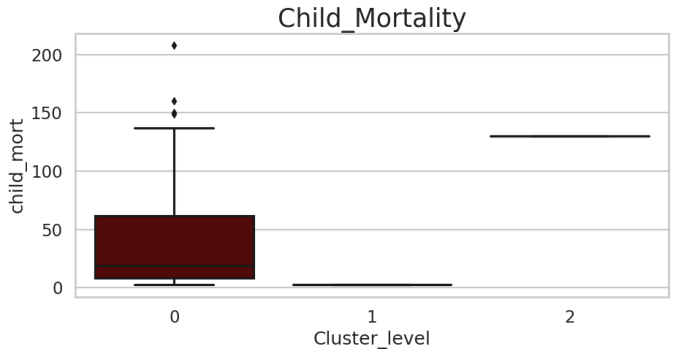
Insight from scatter plot:-

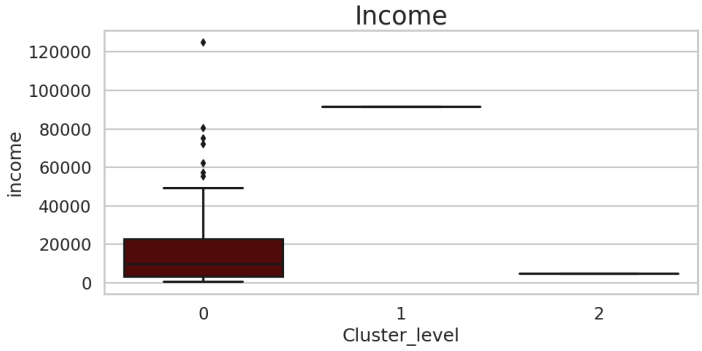
a. The PC1 is in the direction of of least help, we choose PC1 because of variance.
b. The left side of 2.5 value of PC1 need help in aid.

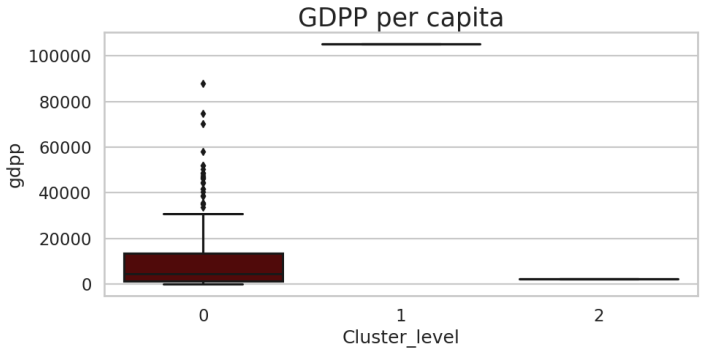
4.Original variables visualization

The boxplots of the variables life expectancy, child mortality, income, and GDPP per capita have been plotted for each cluster.









Insight from plots:-

a. For cluster 0: gdpp and income is the lowest than other clusters, Mortality of children is very high than other clusters.
b. For 1: Behaving normally in all departments excepts for outliers.
c. For 2: gdpp and income is higher than other clusters, and Mortality is less.

Note:-

a. Based on the analysis, the countries belonging to cluster 0 require immediate aid for their socio-economic development. The non-profit organization should focus on these countries for providing aid and support. The countries in cluster 1 and 2 also require aid, but to a lesser extent compared to cluster 0. The non-profit organization can also consider these countries for support.
b. There are 147 countries found from the hierarchical analysis in need of urgent help.

5.Insight(Approach1- with outlier)

K-Means vs Hierarchical Clustering

1. **K-means clustering:**
    a. There are total 147 countries are in this categories
    b. Countries that are having good socio-economic and health factors
    c. Total 2 countries are in this category-Luxemburg and Singapore

2. **Hierarchical clustering:**

    a.Countries with direst need of aid are as:

    1.Total 147 countries are in this category

    2.1 country with good socio-economic and health factors - Luxembourg

Conclusion:

After inspecting both clustering method, we can say that the final countries from K-means clustering as it gives accurate output than hierarchical clustering.I have compared the clusters and visualized from methods and k-means gave precise information than hierarchical clustering.\

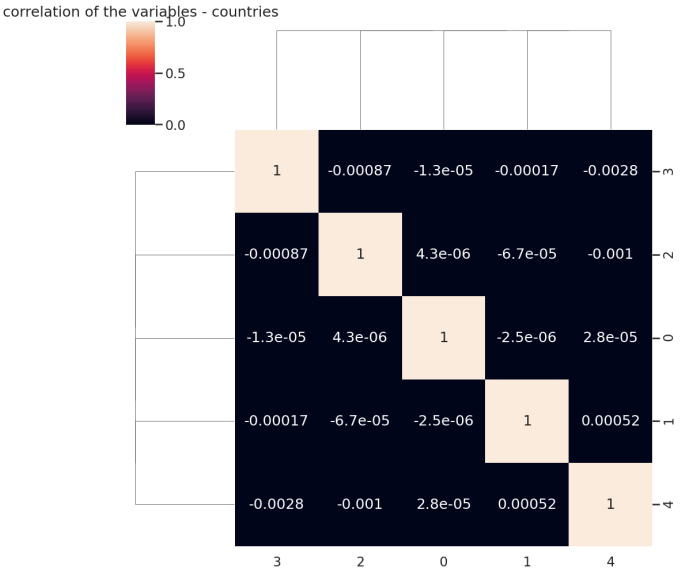
6.Exclude outliers(Approach 2)

K-Means algorithm on the PCA-transformed data. Before clustering, we performs data preprocessing to **remove outliers** using the IQR method.

1.Principal Component Analysis:

- We performed PCA to reduce the dimensions of the data to 5 variables.
- We checked for the correlation between the variables using a heatmap.

Visualization using clustermap:



a. All the correlation are showing in dark color, which means they all are close to 0.
b. This shows that after doing PCA we have removed the multicollinearity.

2.Cluster Analysis:

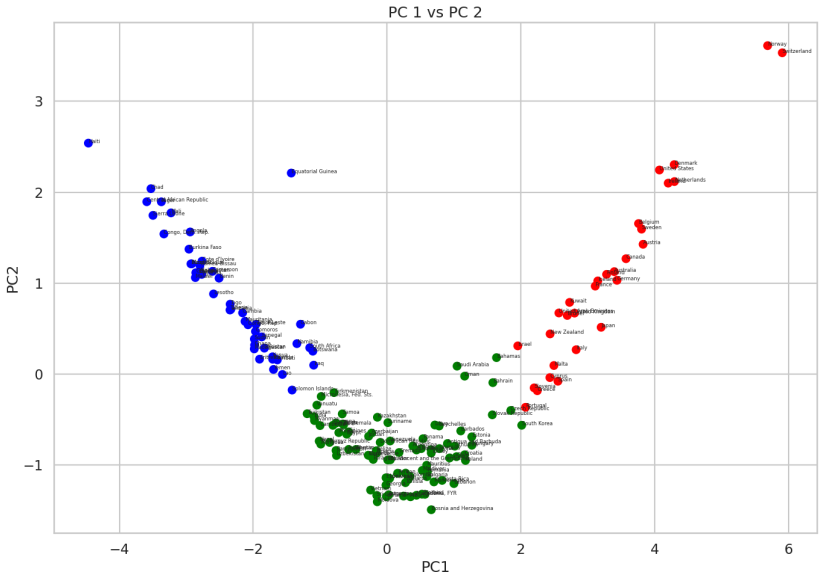
- We used the silhouette score method to determine the optimal number of clusters.

```
Cluster=2 has silhouette score 0.3828606546658335
Cluster=3 has silhouette score 0.46180029764974995
Cluster=4 has silhouette score 0.40528890233903064
Cluster=5 has silhouette score 0.3430974759384591
Cluster=6 has silhouette score 0.3509065823550426
Cluster=7 has silhouette score 0.341630473984728 57
Cluster=8 has silhouette score 0.33109295406111844
Cluster=9 has silhouette score 0.3564593643943365
```

- After seeing silhouette score ,we performed KMeans clustering with 3 clusters.

3.Visualization

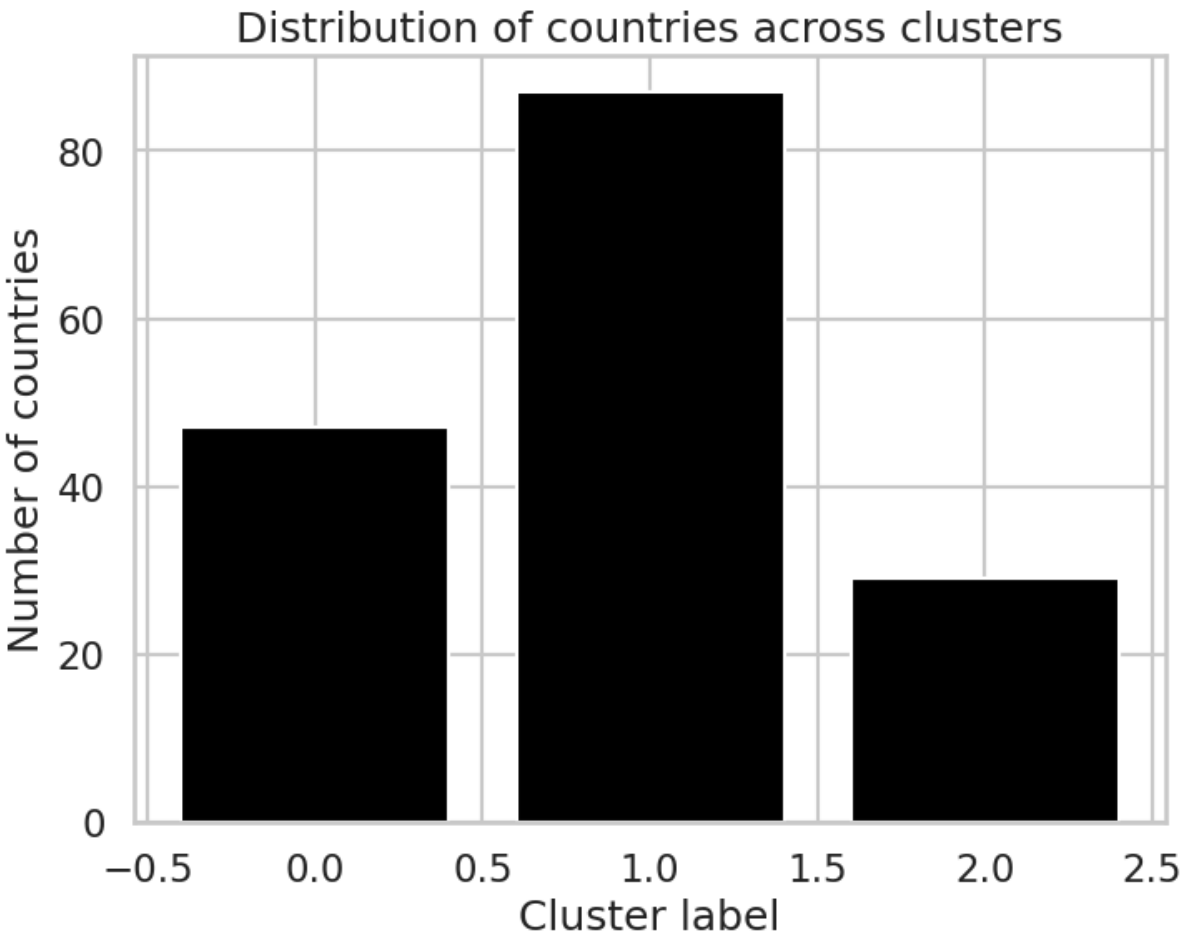
1. PC1 and PC2(Scatter)

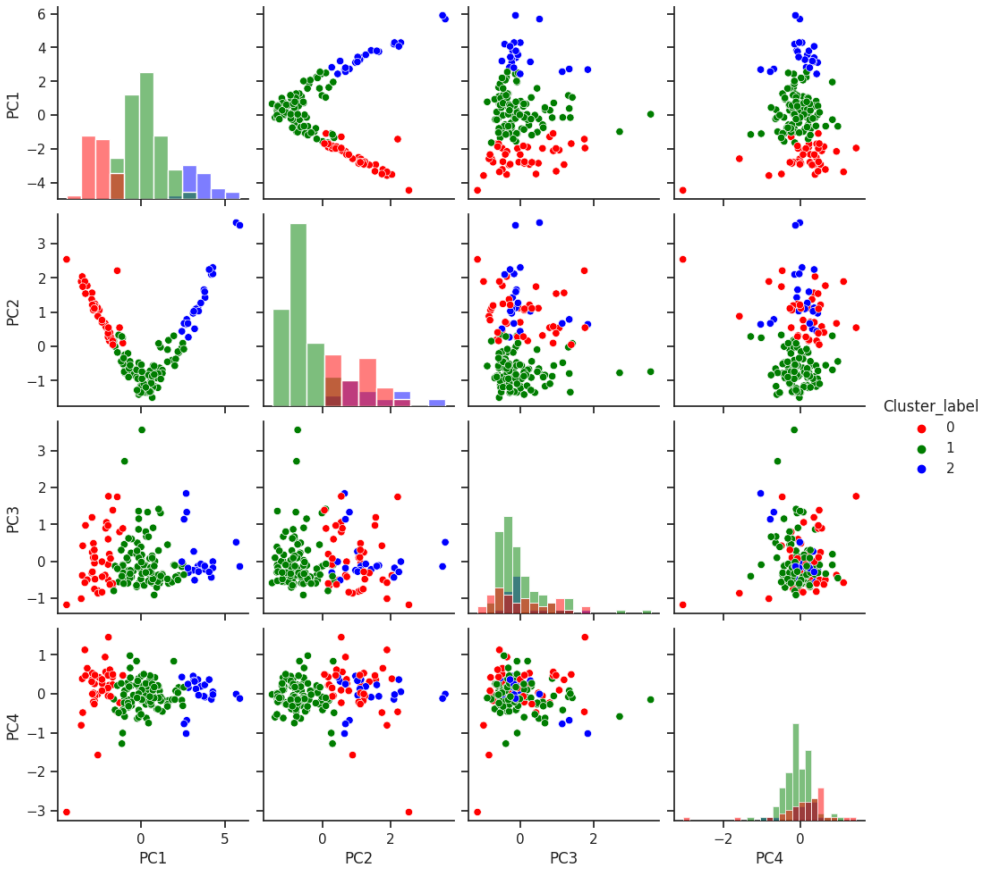


Form graph we concluded that:

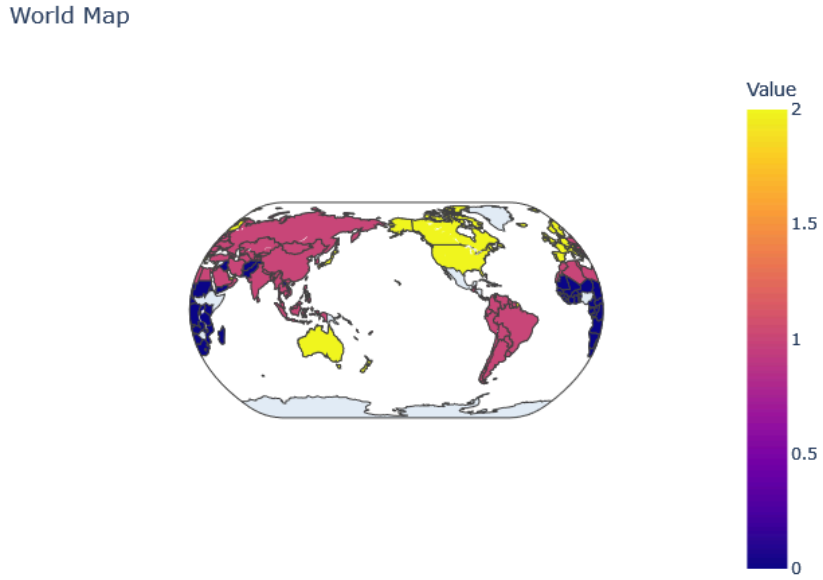
a. We would choose PC1 because it has maximum percentage of variance explained.
b. The Blued color datapoints need urgent help in aid but the “Green” one not required.

2.Bar graph clustering:

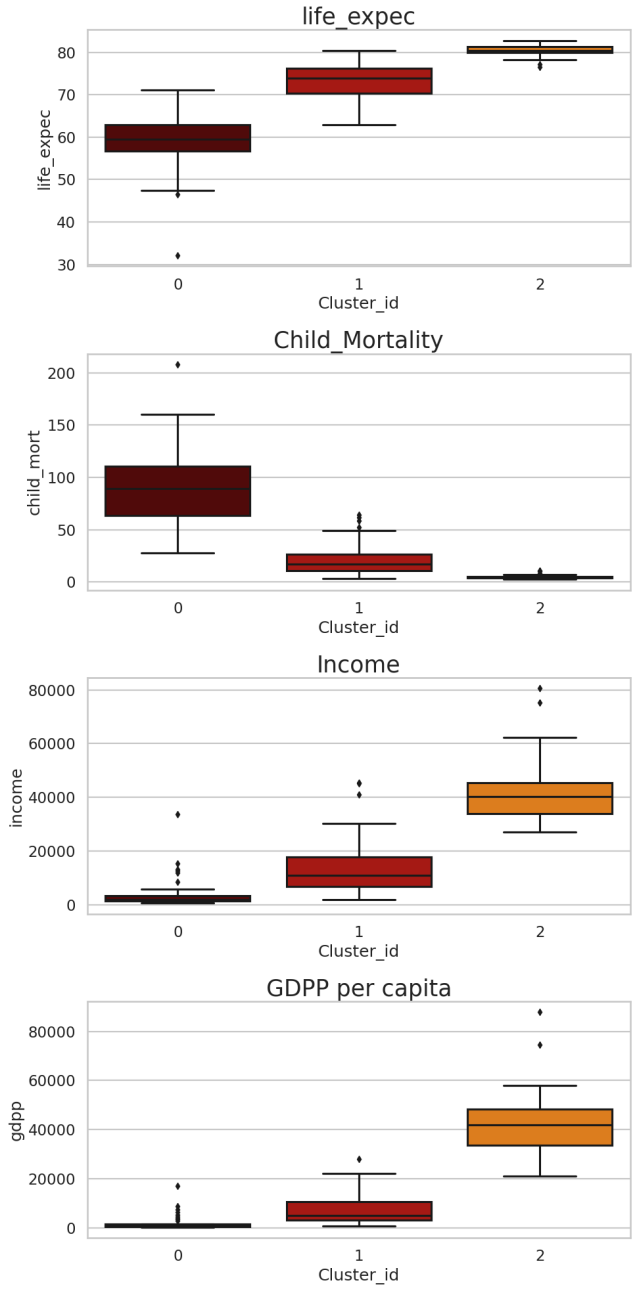




3.Plotly Express to create a world map(based on their cluster):



4.Visualization of original variables(gdpp,income and child mortality)



Insight from the boxplot:

a. For cluster 0: having little higher gdpp and income that cluster 1 and child mortality also acts same.

b. For 1: gdpp and income is higher than other clusters, Mortality of children is very less compared to other countries.
c. For 2: gdpp and income is the lowest than other clusters, Mortality of children is very high than other clusters.

Results:

The silhouette score analysis showed that 3 clusters would be the optimal number of clusters for the dataset. We performed KMeans clustering with 3 clusters and plotted the countries on a world map. The analysis showed that most of the African countries were in cluster 2, which had the lowest values for health and socio-economic factors. On the other hand, the European and American countries were in cluster 0, which had the highest values for health and socio-economic factors.

Conclusion:

1. **We performed a cluster analysis on a dataset of 167 countries based on their health and socio-economic factors.**
2. **There are a total 47 countries from the dataset in need of urgent help as they are having lowest income,high child mortality and low gdp per capita.**
3. **There are a total of 28 countries having good socio-economic and health factors.**
4. **We used KMeans clustering algorithm to cluster the countries into 3 clusters.**
5. **The analysis showed that African countries were in the cluster with the lowest values for health and socio-economic factors, while European and American countries were in the cluster with the highest values for these factors**
6. **This analysis can help HELP organisation and NGOs to understand the differences between countries and develop policies to address the disparities, and provide necessary aid/help.**

7. **Cluster with ClusterID as 0, is the cluster of most backward country.**

<u>Countries on which we require to focus more are:-</u>

Afghanistan', 'Benin', 'Botswana', 'Burkina Faso', 'Burundi','Cameroon', 'Central African Republic', 'Chad', 'Comoros','Congo, Dem. Rep.', "Cote d'Ivoire", 'Eritrea', 'Gabon', 'Gambia','Ghana', 'Guinea', 'Guinea-Bissau', 'Haiti', 'Iraq', 'Kenya','Kiribati', 'Lao', 'Lesotho', 'Liberia', 'Madagascar', 'Malawi','Mali', 'Micronesia, Fed. Sts.', 'Mozambique', 'Namibia', 'Niger','Nigeria', 'Pakistan', 'Rwanda', 'Senegal', 'Sierra Leone','Solomon Islands', 'South Africa', 'Sudan', 'Tajikistan','Tanzania', 'Timor-Leste', 'Togo', 'Uganda', 'Yemen', 'Zambia".