



# Clustering of countries

(To find out which countries are in need of the aid)



# Objective:

HELP International is an international humanitarian NGO that is committed to fighting poverty and providing the people of backward countries with basic amenities and relief during the time of disasters and natural calamities. It runs a lot of operational projects from time to time along with advocacy drives to raise awareness as well as for funding purposes.



After the recent funding programmes, they have been able to raise around \$ 10 million. Now the CEO of the NGO needs to decide how to use this money strategically and effectively. The significant issues that come while making this decision are mostly related to choosing the countries that are in the direst need of aid.

As a Data Analyst we will find out which countries are in the dire need of aid.

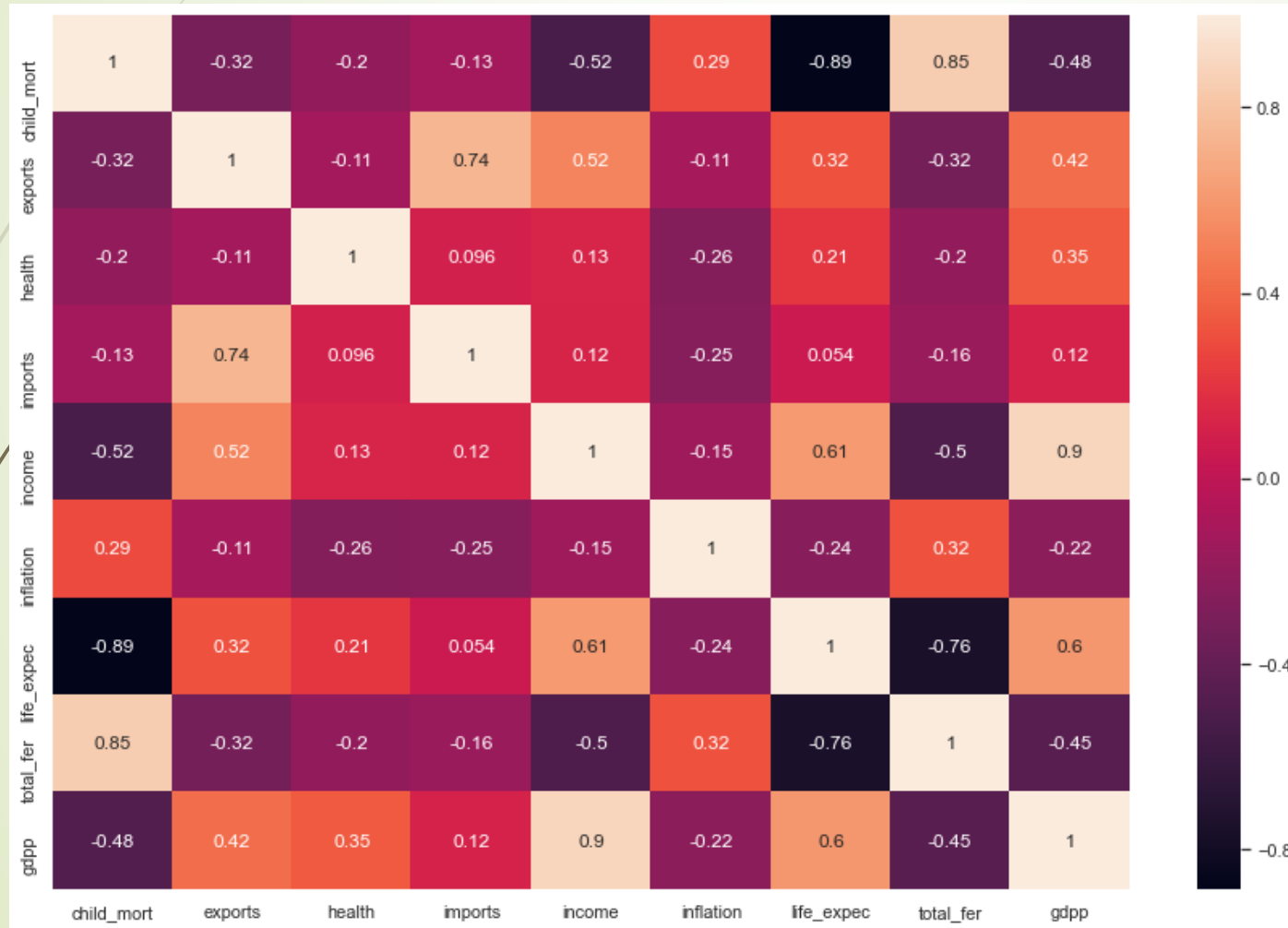


# Analysis Methodology

- Step 1: Reading and Understanding the Data
- Step 2: Data Cleansing
  - Missing Value check
  - Data type check
  - Duplicate check
- Step 3: Data Visualization
  - Heatmap
  - Pairplot
- Step 4: Data Preparation
  - Rescaling
- Step 5: PCA Application
  - Principal Components Selection
  - Outlier Analysis and Treatment

- 
- 
- Step 6: Hopkins Statistics Test
    - Hopkins Score Calculation
  - Step 7: Model Building
    - K-means Clustering
    - Elbow Curve
    - Silhouette Analysis
    - Hierarchical Clustering
  - Step 8: Final Analysis
    - Final Country list Preparation

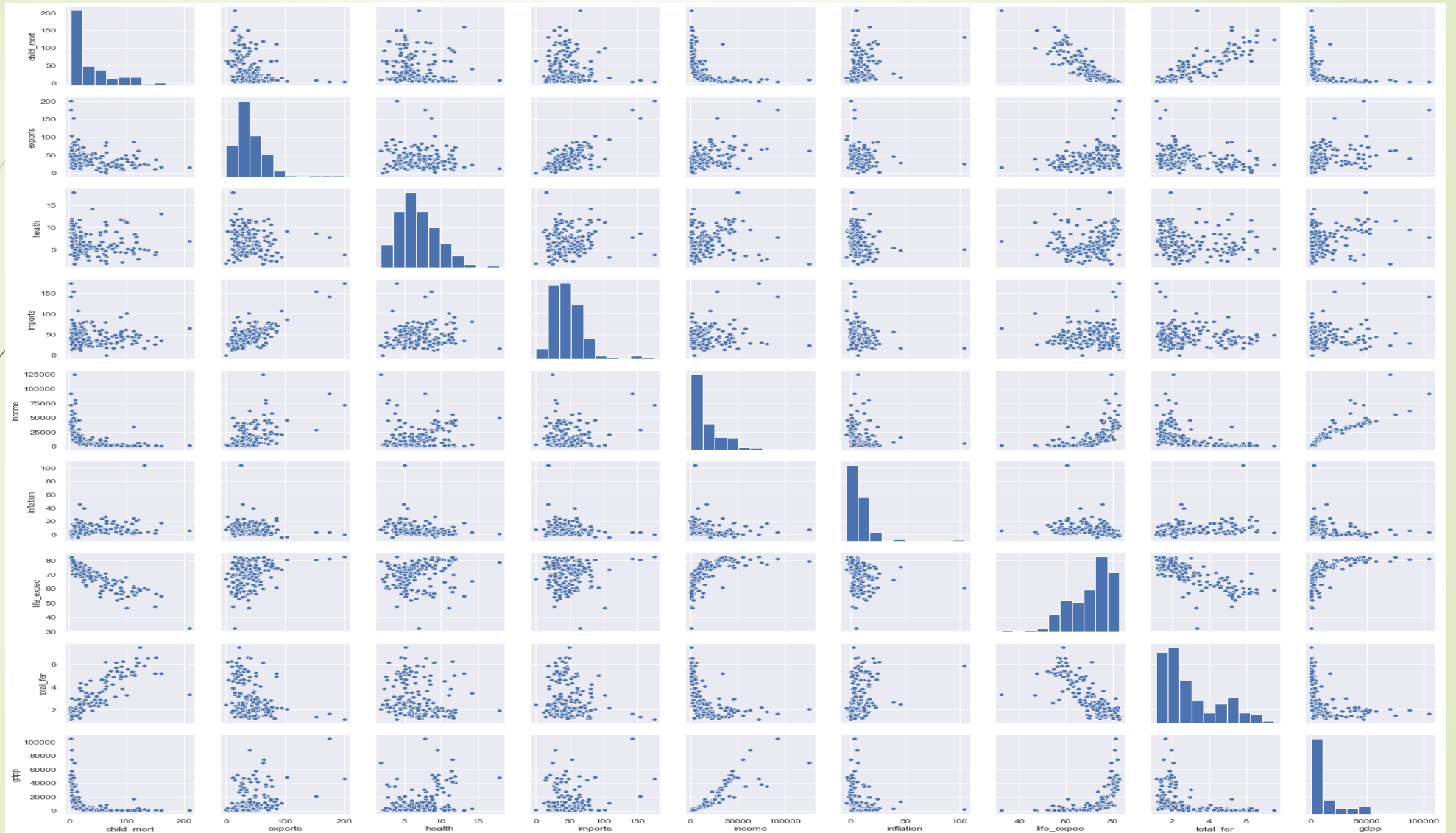
# Heatmap Correlation



Inference:

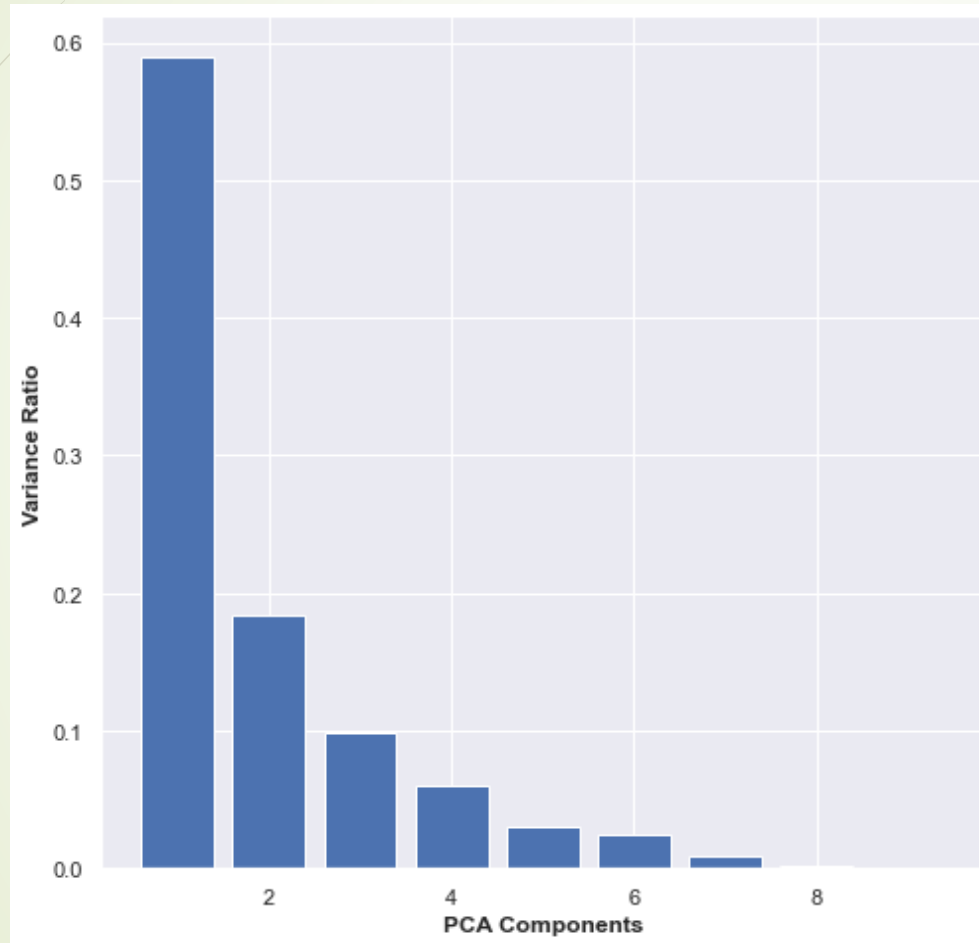
- child mortality and life expectancy are highly correlated.
- Child mortality and total fertility are highly correlated.
- imports and exports are highly correlated.
- Life expectancy and total fertility are highly correlated.

# Pair plot Visualization of data





# Principal Component Analysis

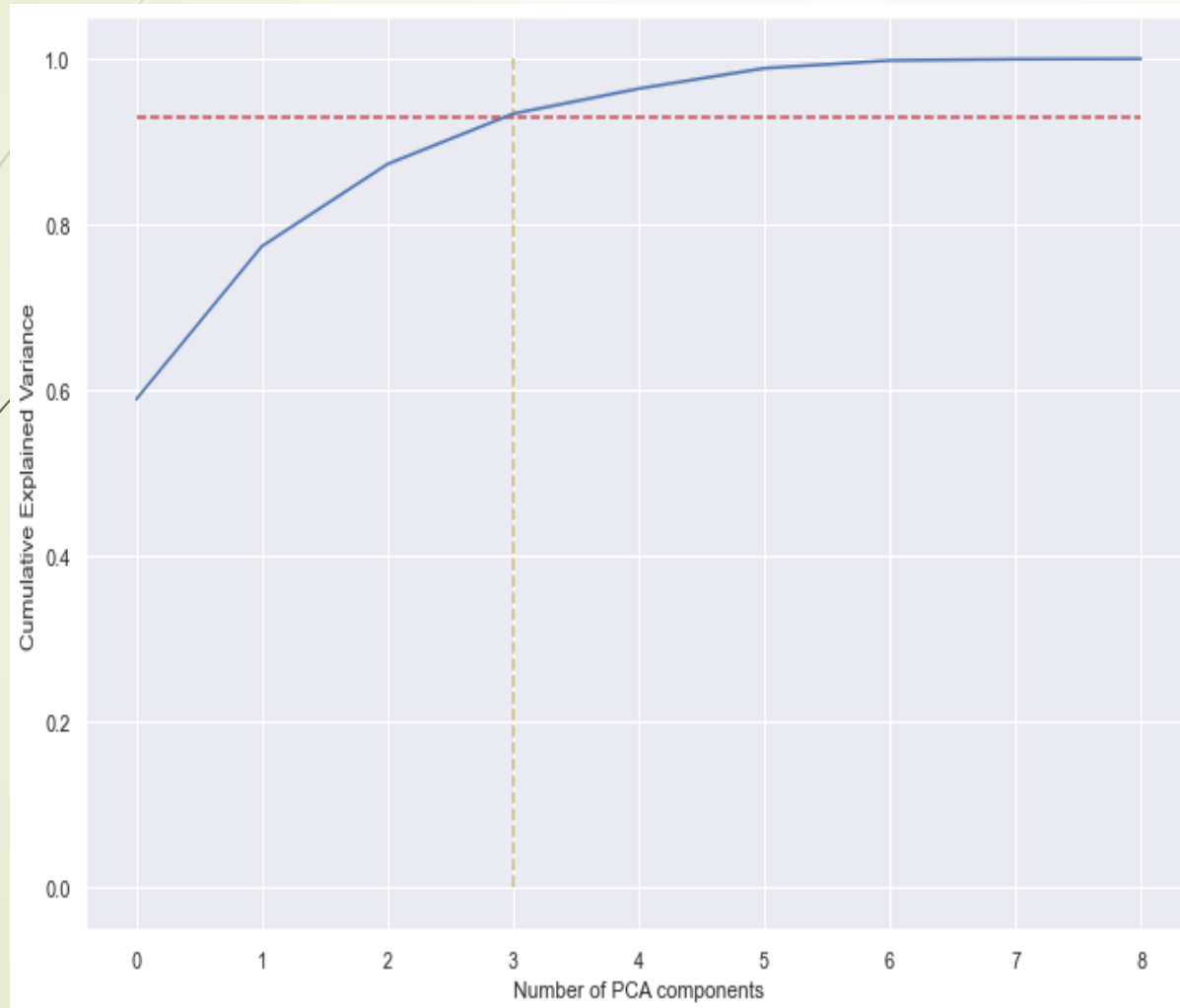


Variance Ratio Bar plot for each PCA components

Inference:

For first component variance is almost 60%.  
For second component variance is almost 20%.

# Principal Component Analysis

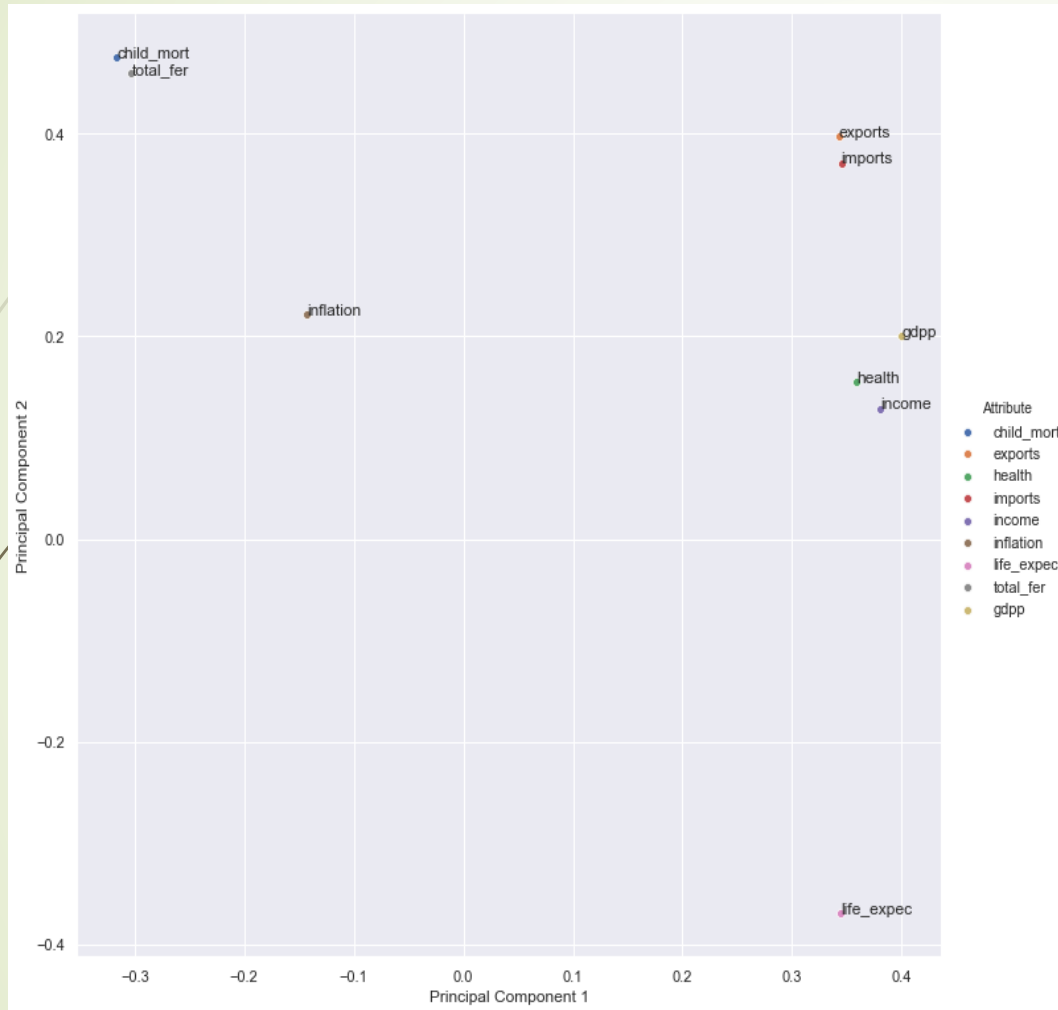


Scree Plot for cumulative variance against number of PCA Components.

**Inference** : more than 90% variance is explained by the first 3 principal components



# Principal Component Analysis



PC1 and PC3:

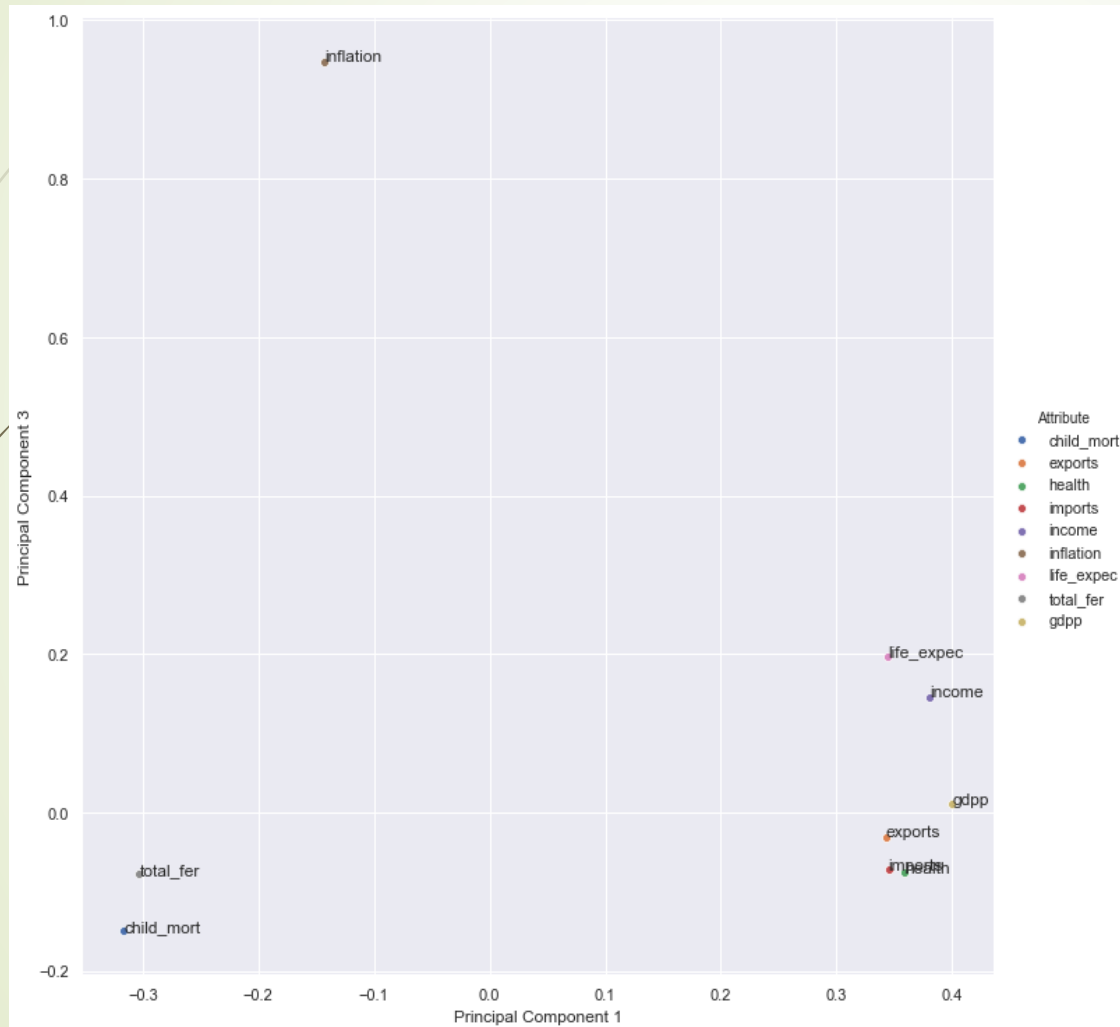
life expectancy, income, gdpp and health are explained by PC1.

imports and exports are explained by PC1 and PC2.

child mortality and total fertility are explained by PC2.

inflation is neither explained by PC1 nor with PC2.

# Principal Component Analysis

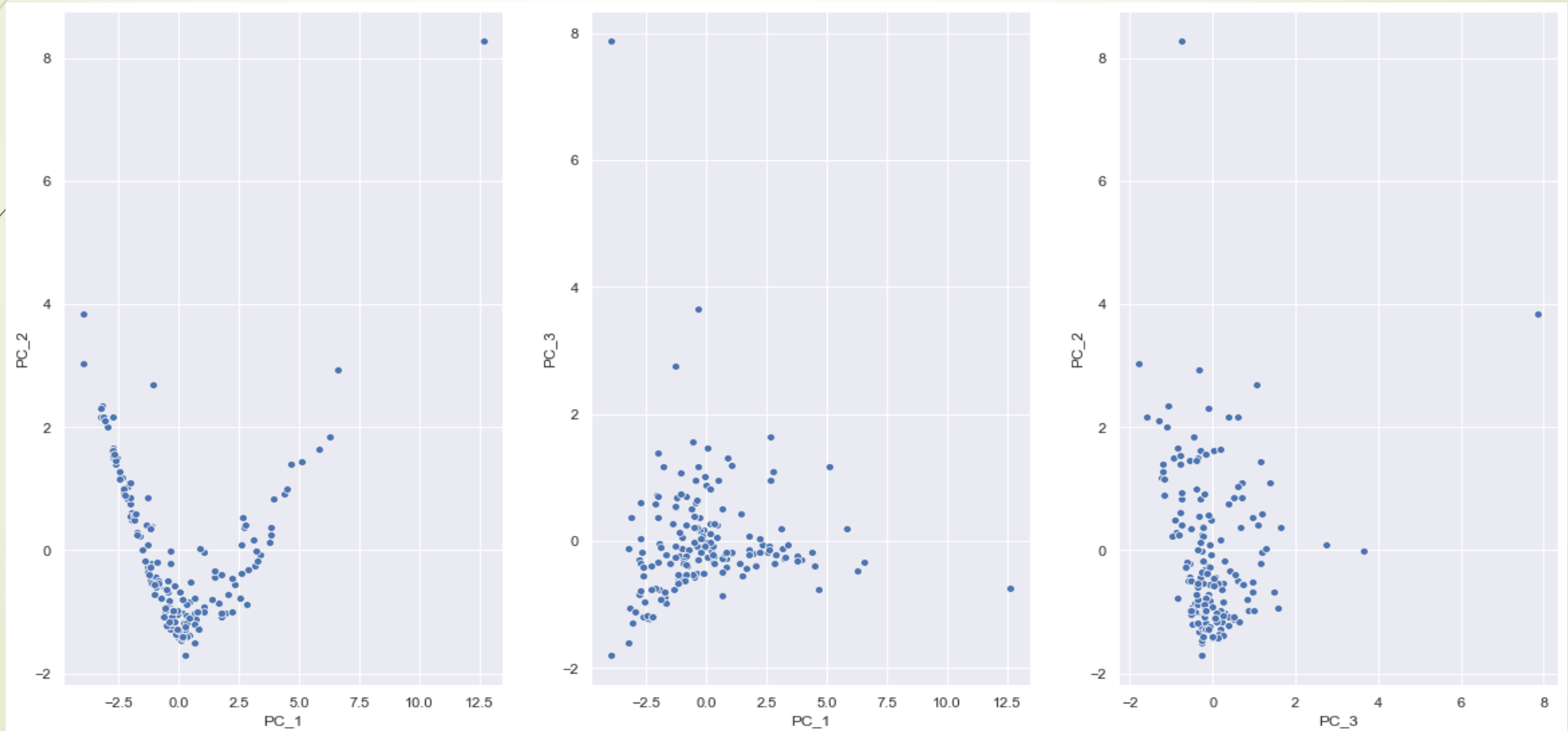


PC1 and PC3:

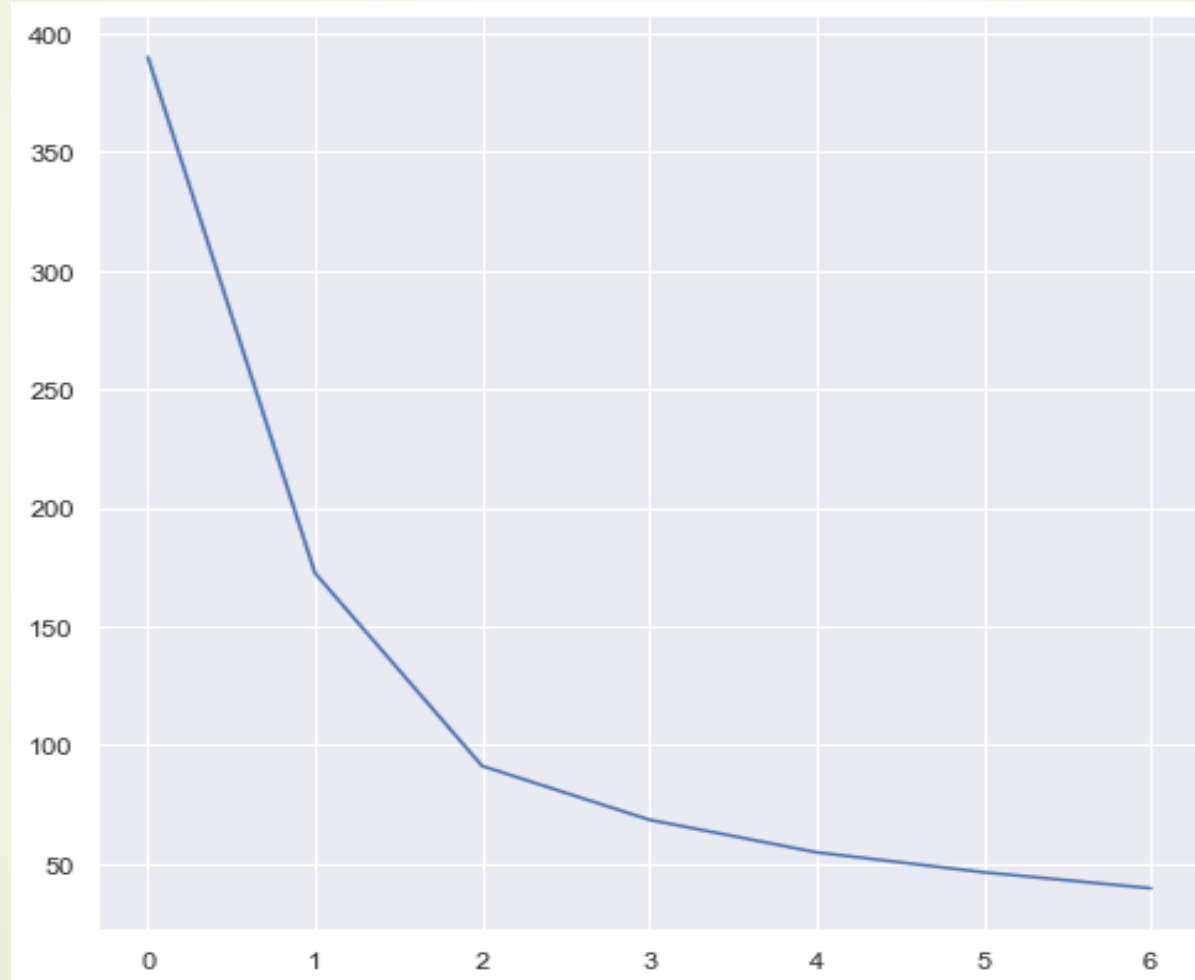
Inflation is  
explained by PC3

# Principal Components Analysis

Spread of data across various PCA components



# K-Means Clustering (Elbow Curve)

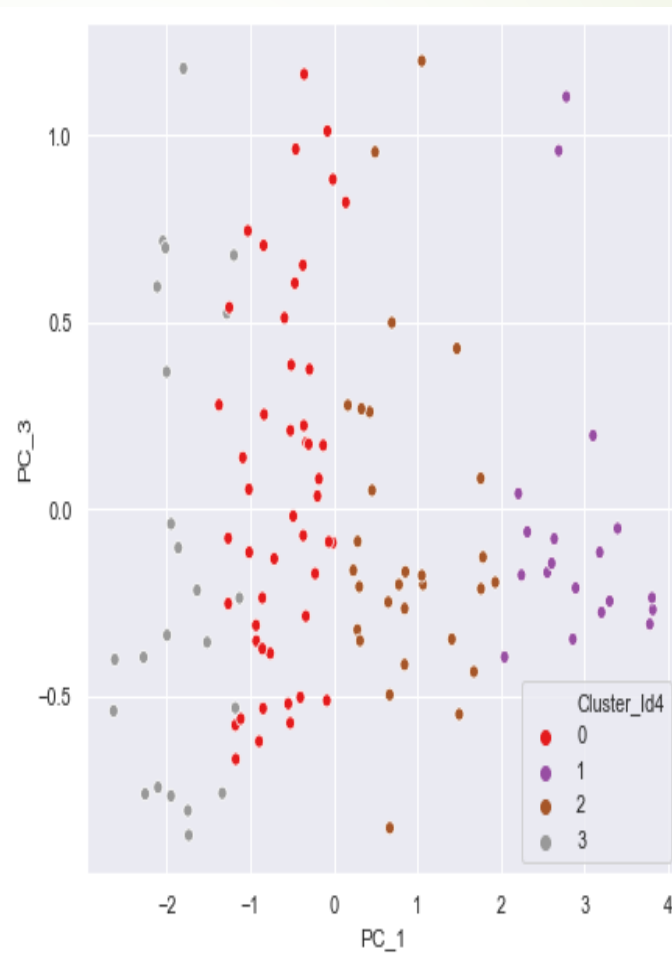
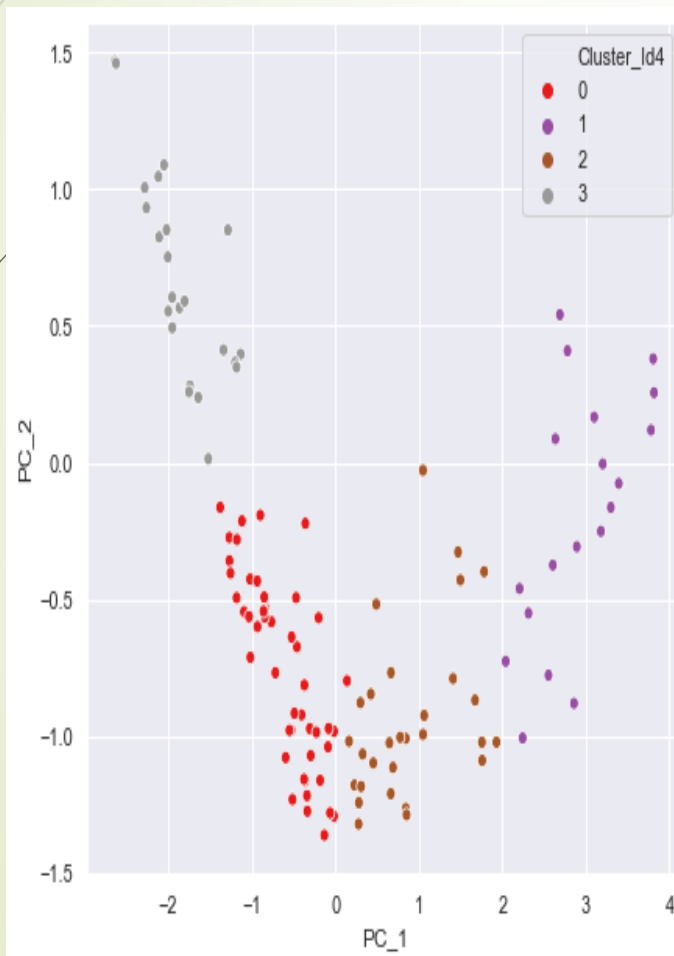


Number of Clusters : 4 or 5

# Silhouette Analysis

- Silhouette Score :
- For n\_clusters=2, the silhouette score is 0.4873400103541441
- For n\_clusters=3, the silhouette score is 0.4639771456218248
- For n\_clusters=4, the silhouette score is 0.3987356568367148
- For n\_clusters=5, the silhouette score is 0.36170980333920066
- For n\_clusters=6, the silhouette score is 0.36603716544306125
- For n\_clusters=7, the silhouette score is 0.3703642483431638
- For n\_clusters=8, the silhouette score is 0.3752370154601887

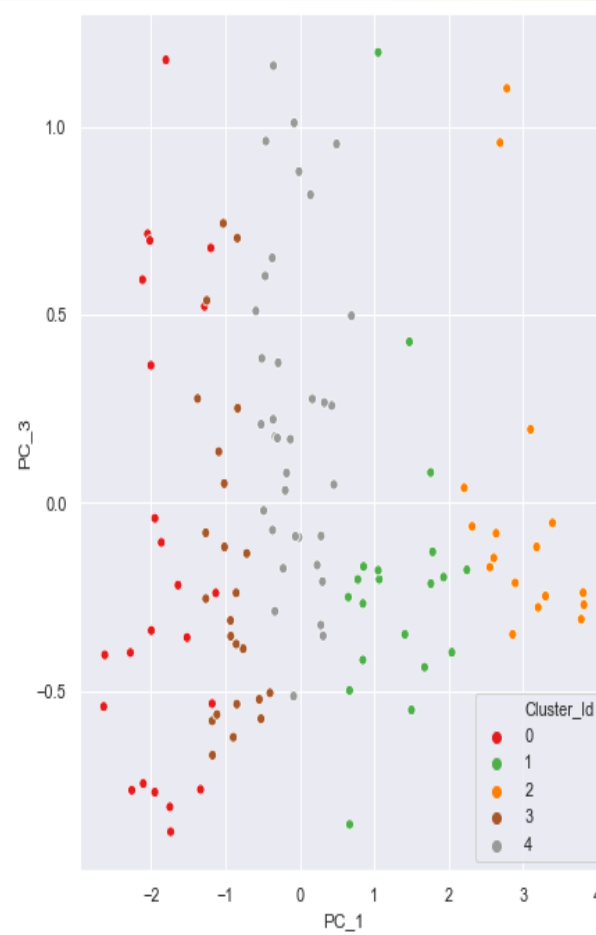
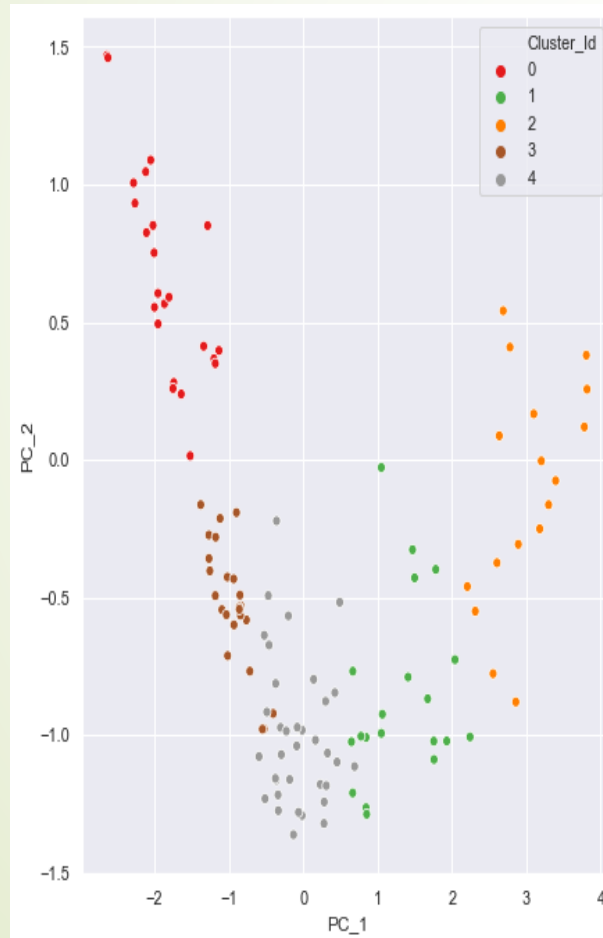
# Visualization of data spread with number of clusters = 4



Inference:

lot of intra-distance between the cluster elements.

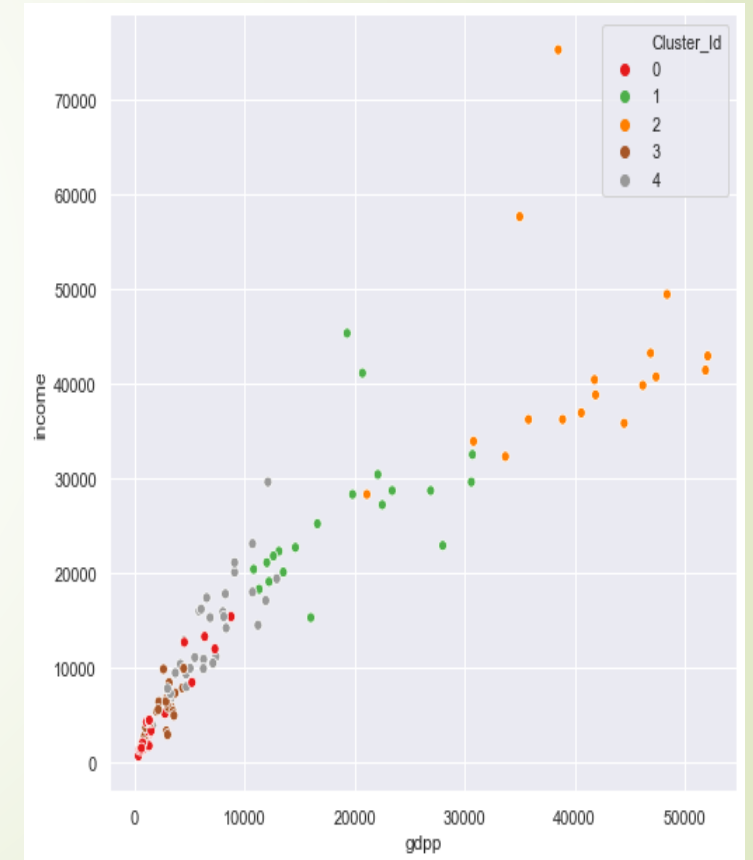
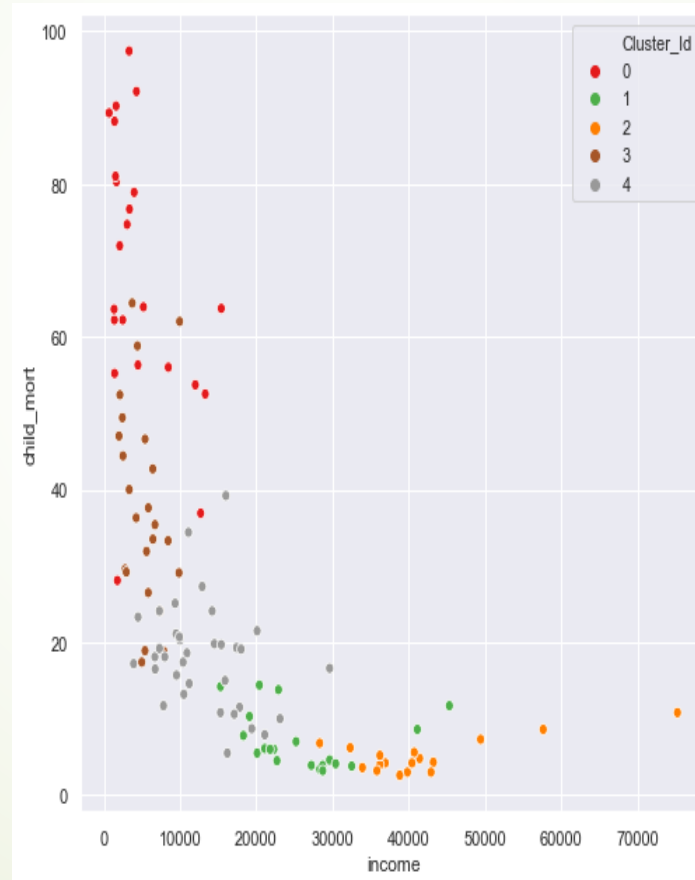
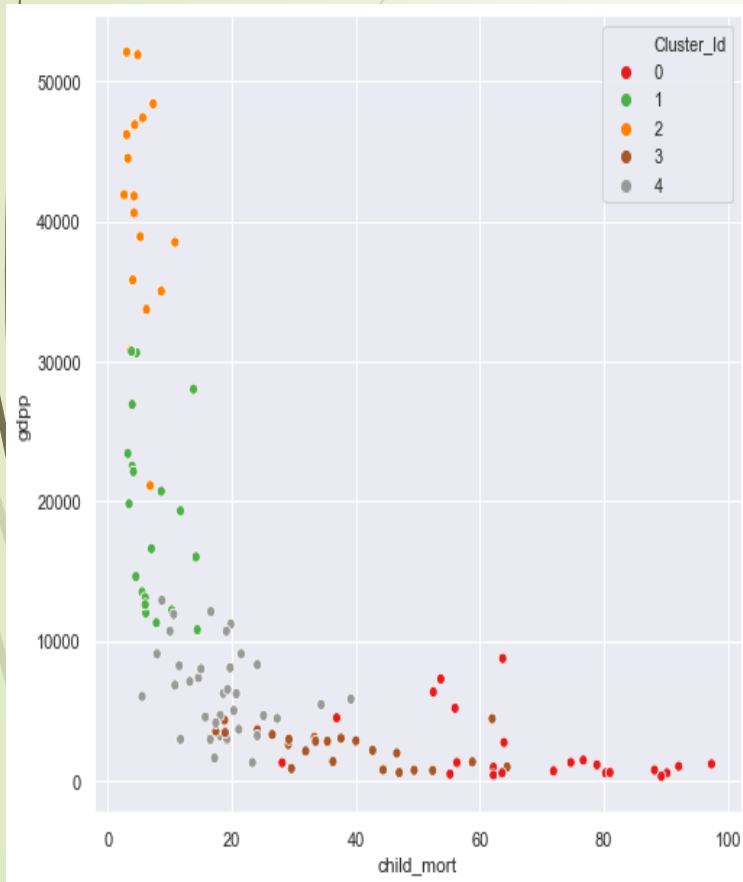
# Visualization of data spread with number of clusters = 5



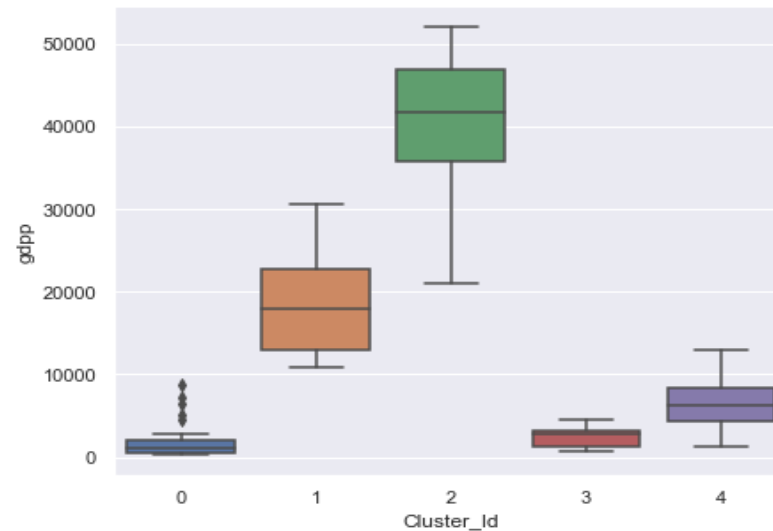
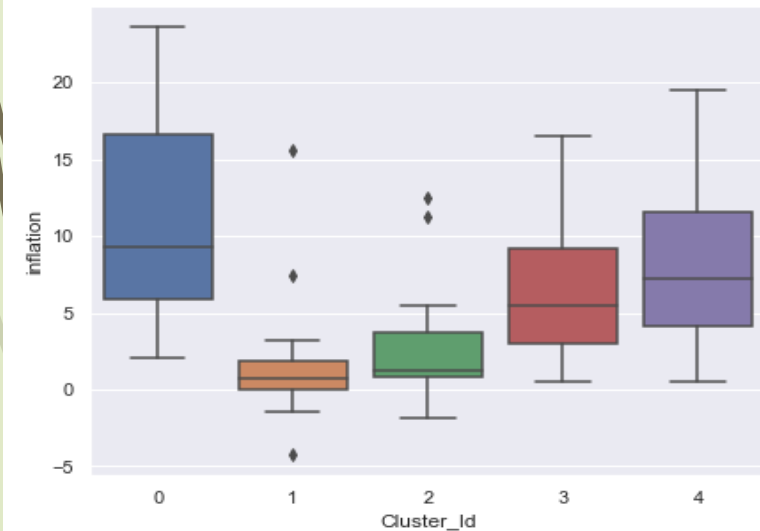
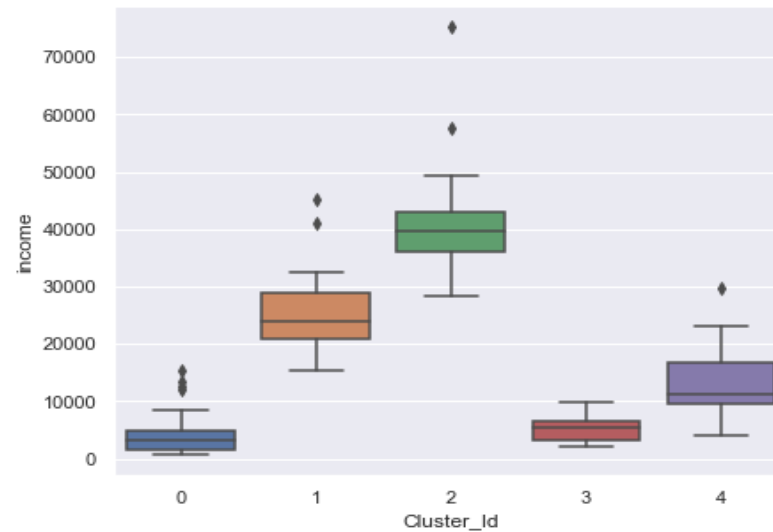
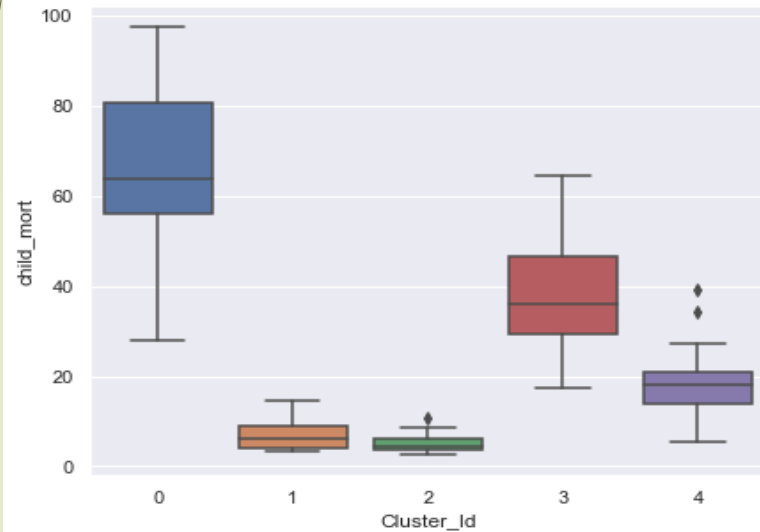
Same issue as number of clusters = 4 so, proceeding with number of clusters = 5



# Scatter Plot for Child Mortality, Income and GDP



# Boxplot for child mortality, income, inflation and gdp



Inference:

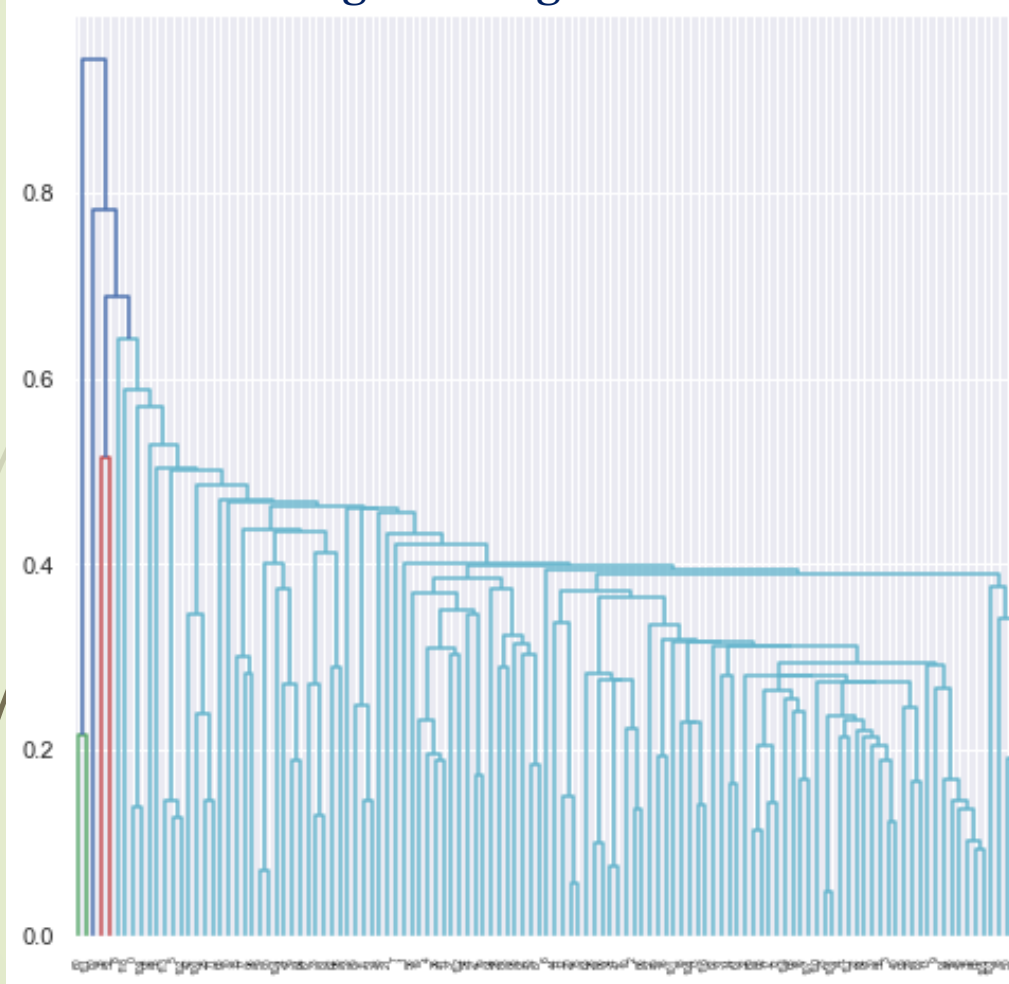
Child Mortality is high for Cluster 0 and Cluster 3.

Income per capita and gdp seems low for countries in clusters 0 and 3.

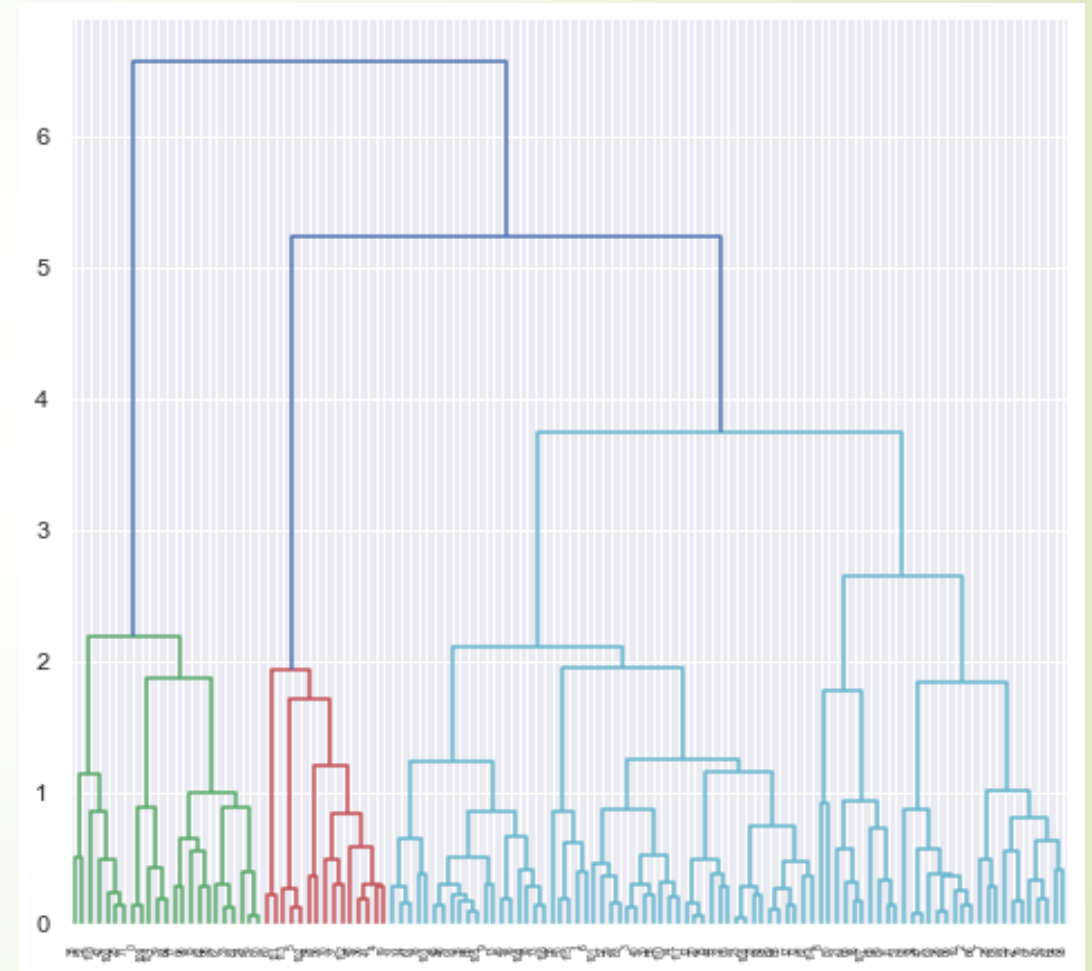
Hence Countries in clusters 0 and 3 need help and aid.

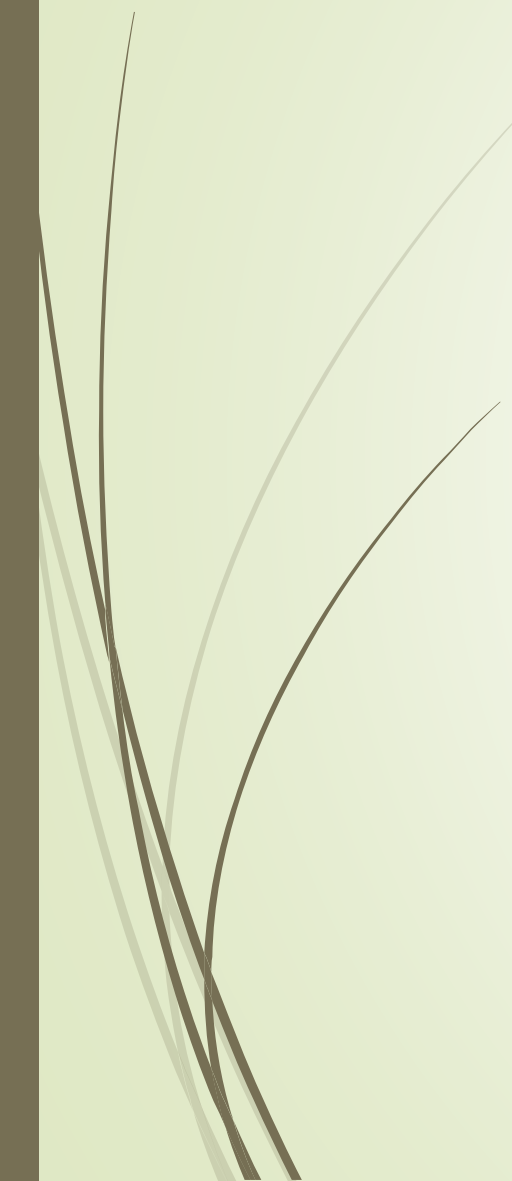

# Hierarchical Clustering

Single Linkage



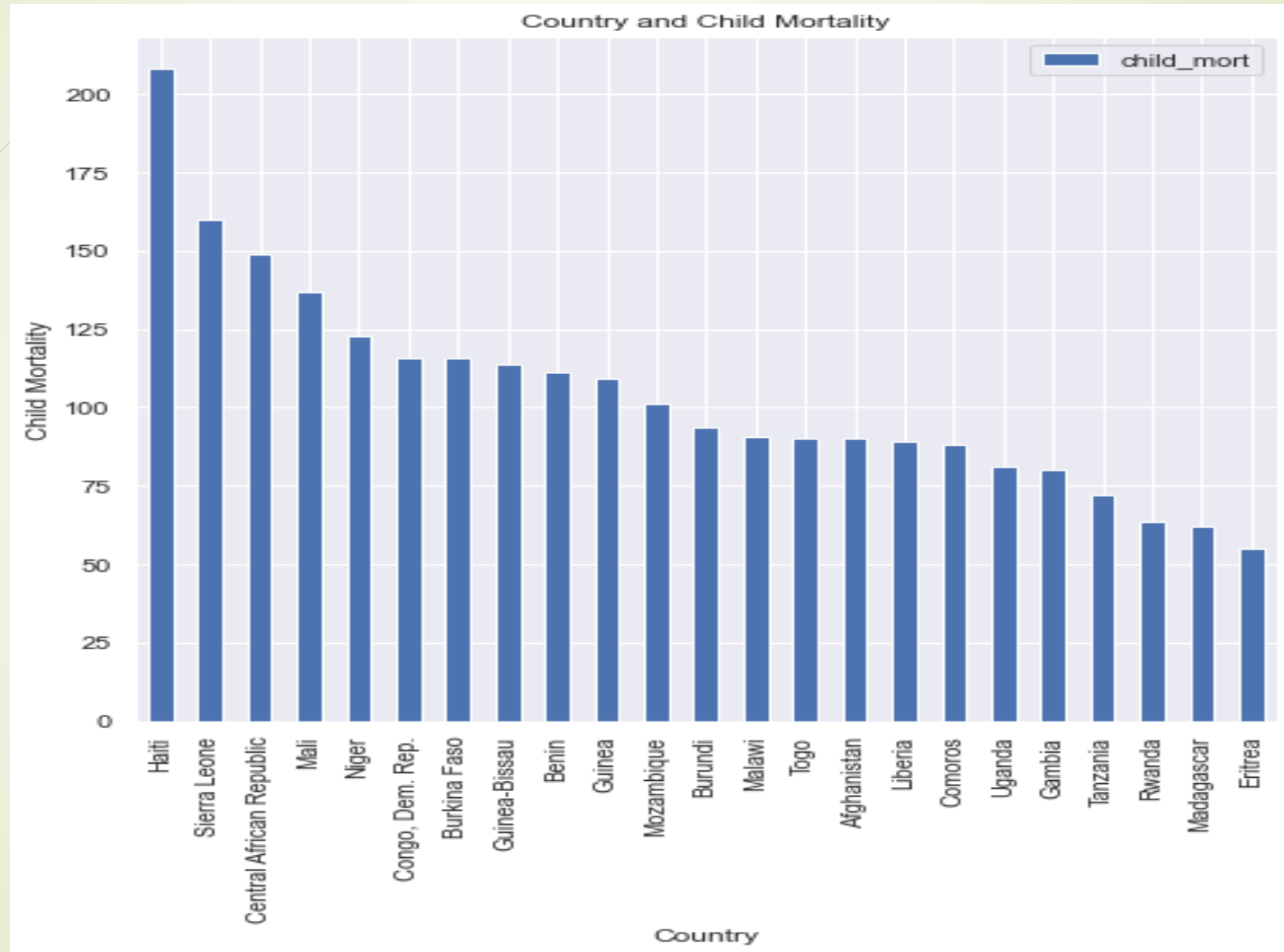
Complete Linkage



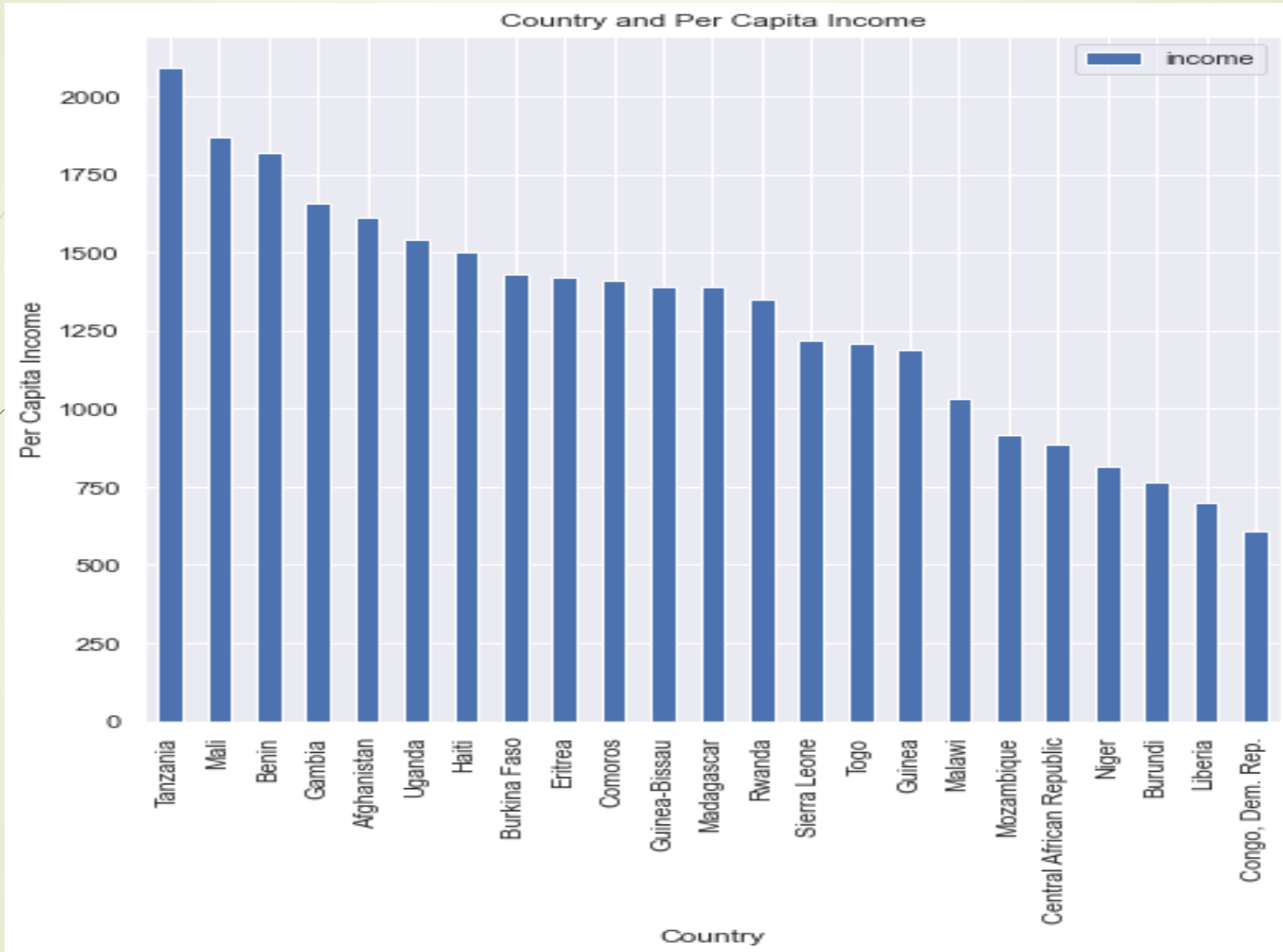


After performing the same process of scatter plot visualisation in the case of hierarchical clustering, It seems that K-means provided better information than Hierarchical Clustering. So, we proceed for the final analysis with K-Means. We merged the countries of cluster 0 and cluster 3 for final analysis.

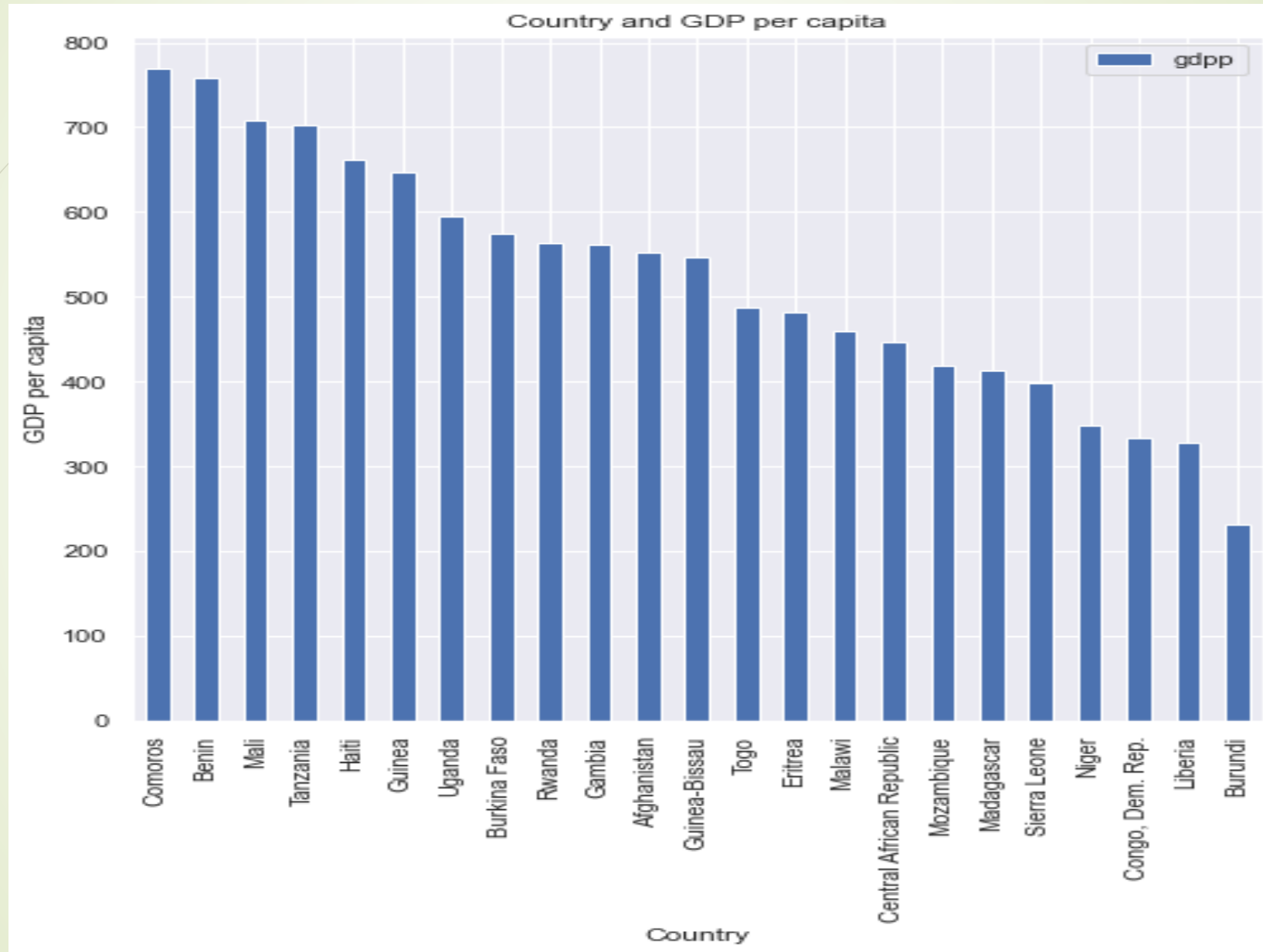
# Country vs Child Mortality



# Country vs Income per capita



# Country vs GDP per capita





# Final List of Countries which are in the need of aid.

- Afghanistan
- Benin
- Burkina Faso
- Burundi
- Central African Republic
- Comoros
- Congo, Dem. Rep.
- Eritrea
- Gambia
- Guinea
- Guinea-Bissau
- Haiti
- Liberia
- Madagascar
- Malawi
- Mali
- Mozambique
- Niger
- Rwanda
- Sierra Leone
- Tanzania
- Togo
- Uganda

