Clustering PCA Assignment

By: Hemant Sawakare

Date: 11/01/2019

Roll Number - DDS1930106

Problem Statement

- HELP International has raised $10M for fighting poverty and providing aid to backward countries.
- Group countries based on various socio-economics factors.
- Identify countries which needs focus based on socio-economic factors i.e. in the direst need of aid.
- Report back at least 5 countries which are in direst need of aid from the analysis work that you perform.

Data Exploration

- **Datasets:**
  - Country data for 167 countries.
  - Data dictionary

Analysis of Country Data

- Socio-Economic data for 167 countries present.
- There are no missing values in data.
- There are countries with high income, high gdpp and high child mortality.
- These variations in different factors will help us group countries into various clusters.
- At the end we will use child mortality and Life expectancy to determine 5 countries needing aid.

Analysis Approach

- **Read and Analyze data**
  - Read data to pandas dataframe
  - Analyze for missing value and outliers
- **Prepare Data**
  - Standardize data so that they are on same scale.
- **Find Principal Components**
  - Determine optimal number of principal components using Scree plot
  - Create a PCA dataset based on principal components
  - Verify there is no multicollinearity in PCA modified dataset using heatmap
- **Hopkins Score**
  - Ensure that data is good for clustering.

Analysis Approach

- **Clustering using kmeans**
  - Determine optimal number of clusters using elbow curve
  - Determine/validate optimal number of clusters using silhouette score
  - Build model using optimal k (clusters)
  - Visualize clusters for income, gdpp and child mortality
- **Clustering using Hierarchical clustering**
  - Use single linkage to form dendrograms
  - Use complete linkage to form dendrograms
  - Form clusters using “cut_tree”.
  - Visualize clusters for income, gdpp and child mortality.
- **Analysis of clusters and Conclusion/Suggestion**
  - Review plots to determine characteristics of clusters

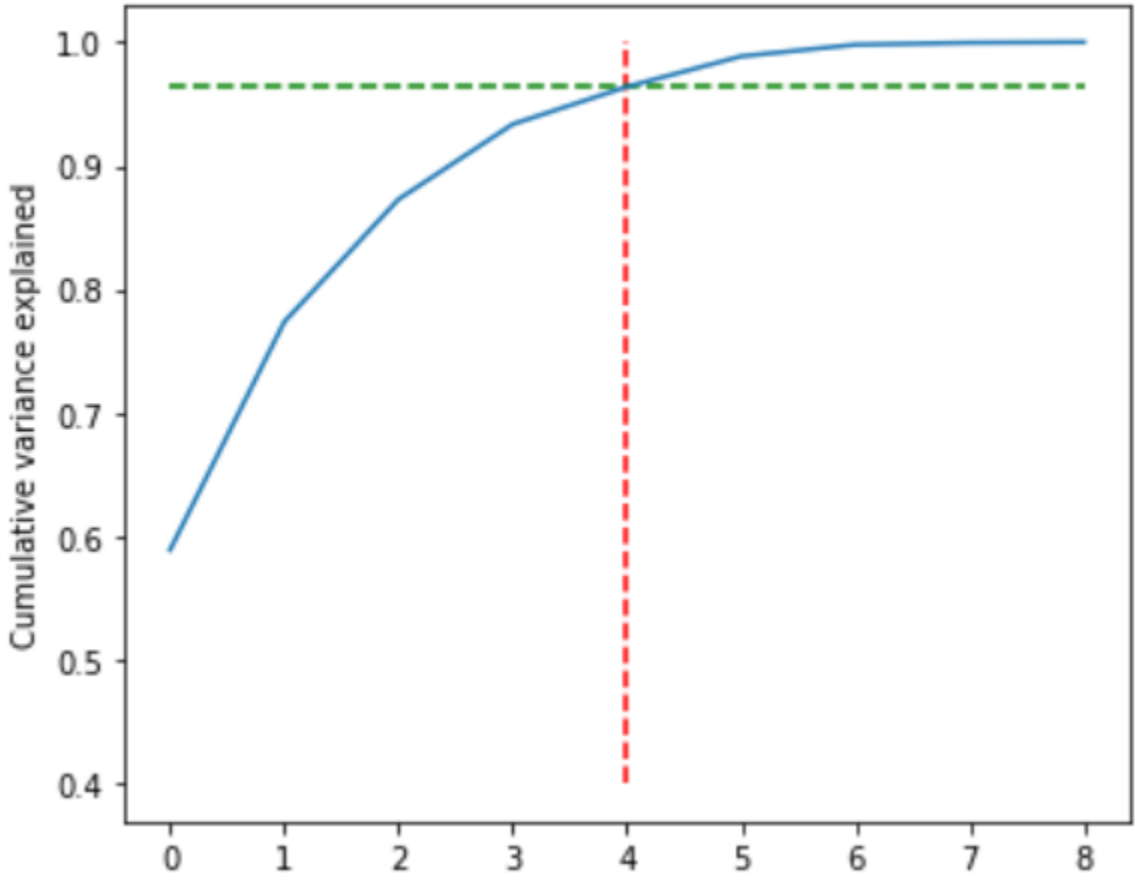
PCA on dataset

Determine Principal Components

- Use scaled data to determining Principal Components (PCs)
- Build PCs using sklearn package PCA.

Review variance explained by Principal Components using “scree plot”

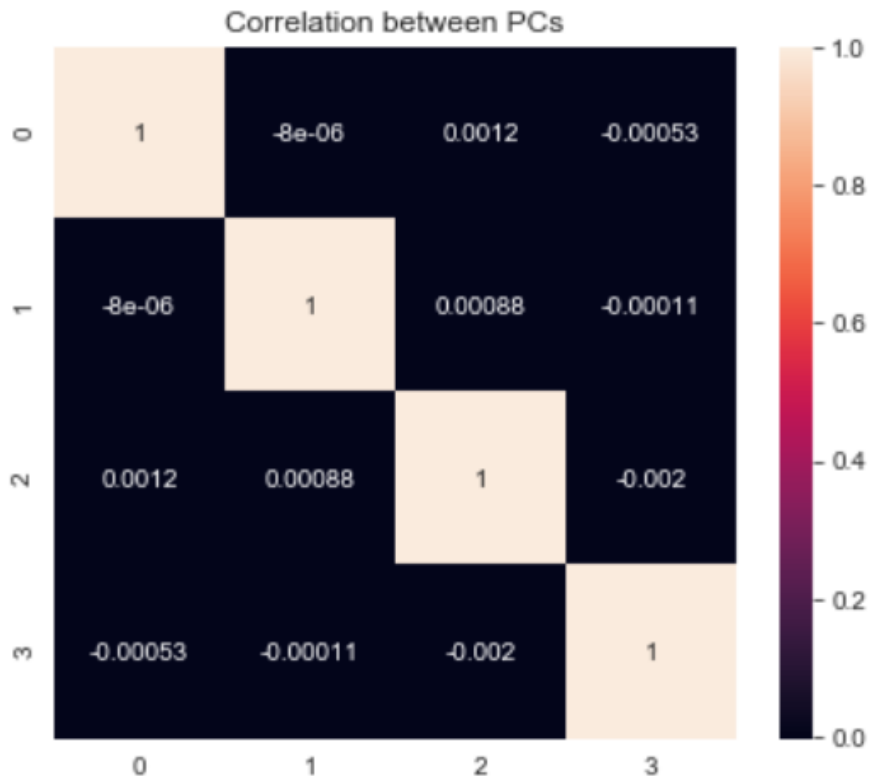
- Check cumulative probability of variance explained by PCs using scree plot.
- As evident from scree plot 95% of variance is explained by 5 PCs.
- Convert data to new PCs using PCA package.



PCA on dataset

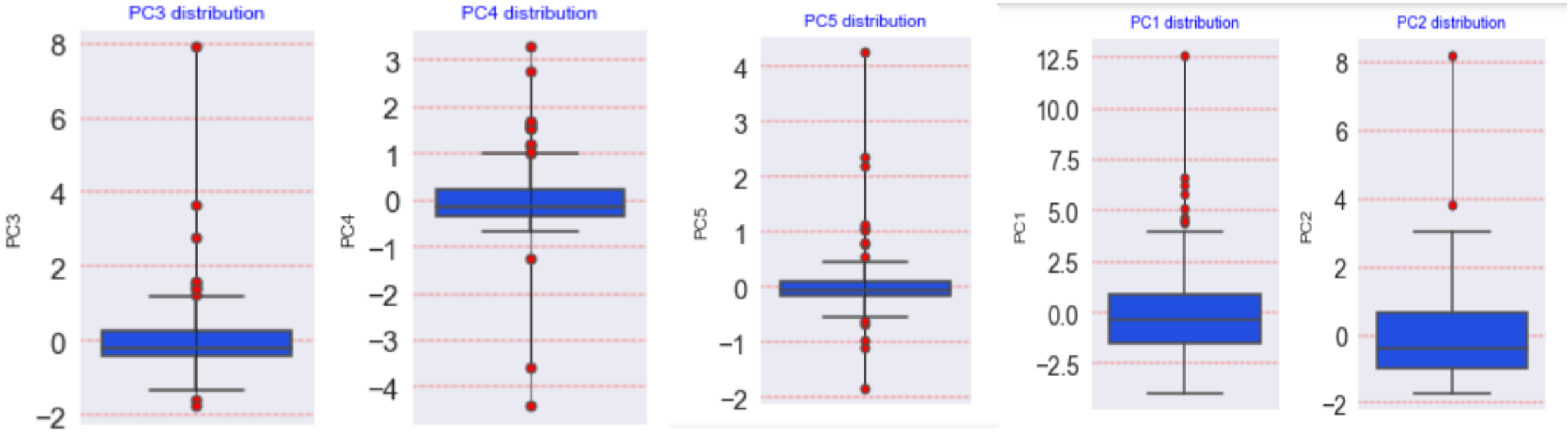
Create dataset with PCs from original country dataset

- Create new dataframe with respect to new PCs.
- New dataset has 5 features and hence have reached dimensional reduction without loss of significant explanatory power.
- **Verify multicollinearity**
  - As evident from heatmap there is no collinearity (nearly 0) between PCs. This will help in building a better model.

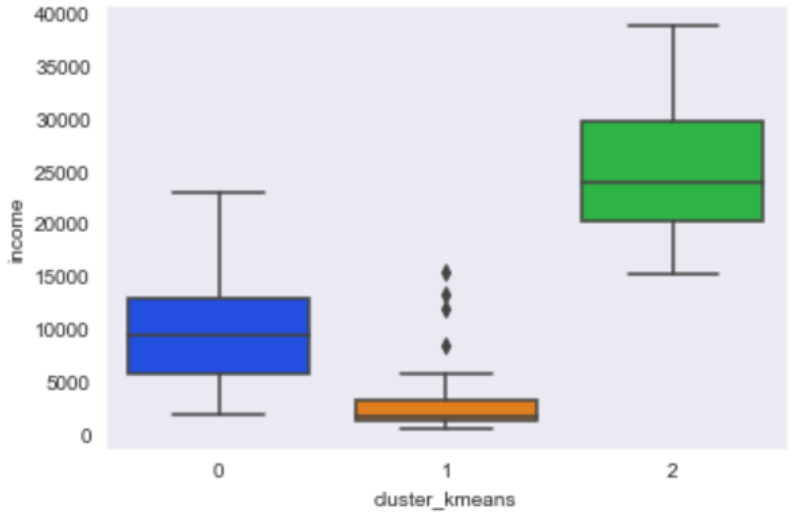


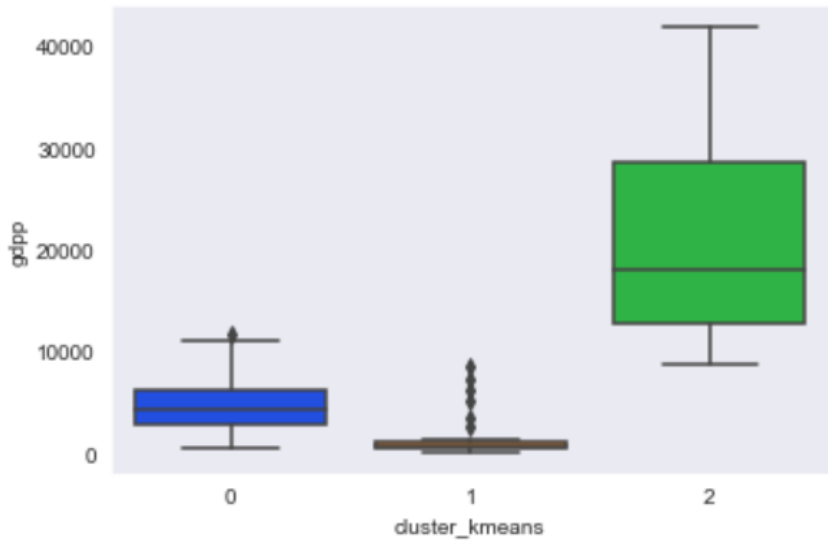
Outlier Treatment

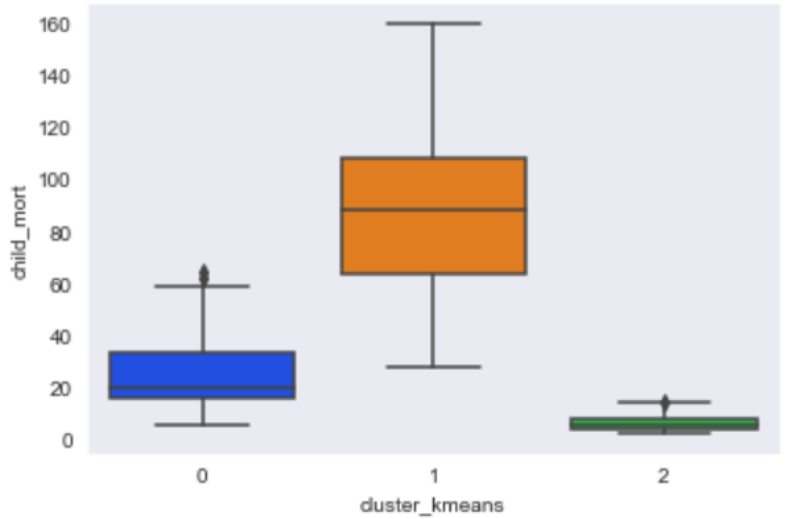
- Outliers create issues for clustering and hence we will be removing them from analysis.
- We will use statistical method for removing the outliers.



Final Clusters - Data Visualization







We will use clusters formed using kmeans clustering.

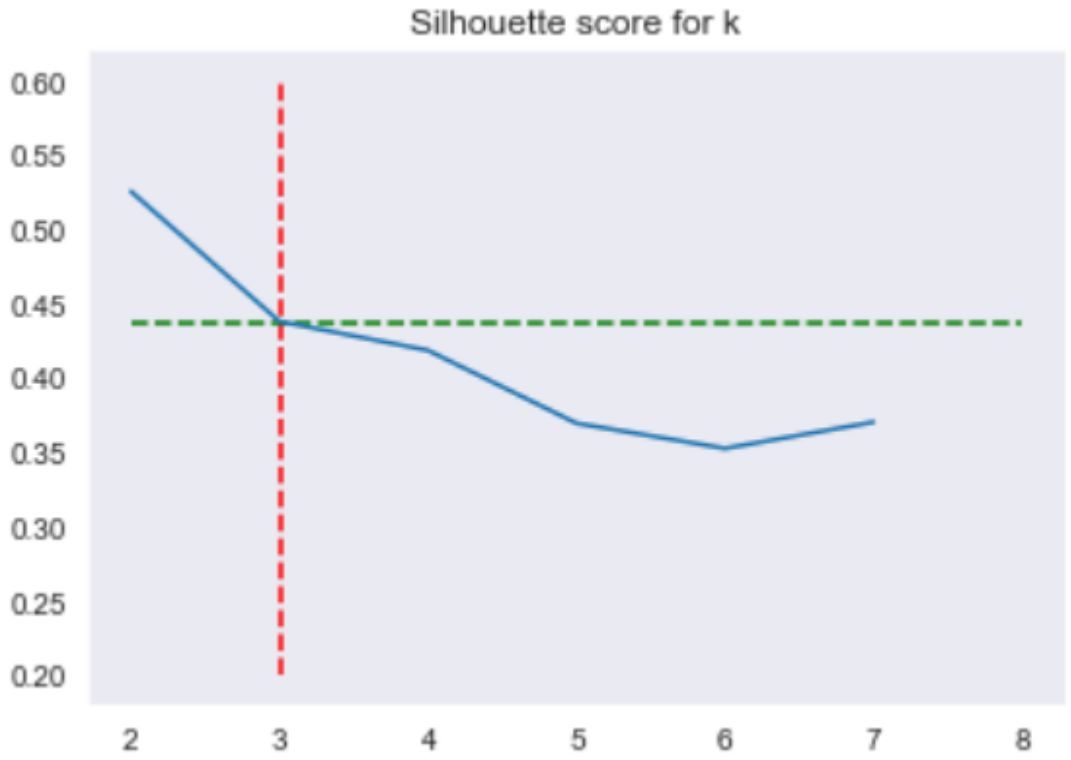
Observations:

- Cluster 0 - Medium gdpp, Medium income, Medium child_mort
- Cluster 1 - Lowest gdpp, Lowest income, Highest child_mort
- Cluster 2 - Highest gdpp, Highest income, Lowest child_mort

Hopkins Score

- Hopkins score indicate clustering tendacy of dataset.
  - Score < 0.3 – Regularly spaced
  - Score < 0.7 – Randomly spaced
  - Score > 0.7 – High tendency of clustering

- For our data, the Hopkins score if > 0.80 and hence we can cluster our data.

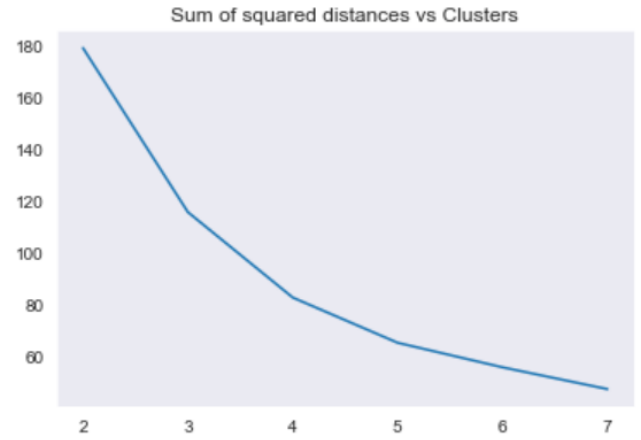


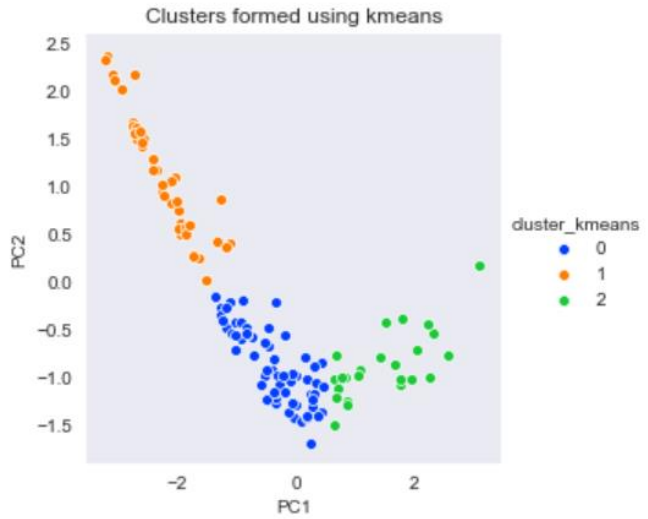
Clustering on PCA dataset (kmeans)

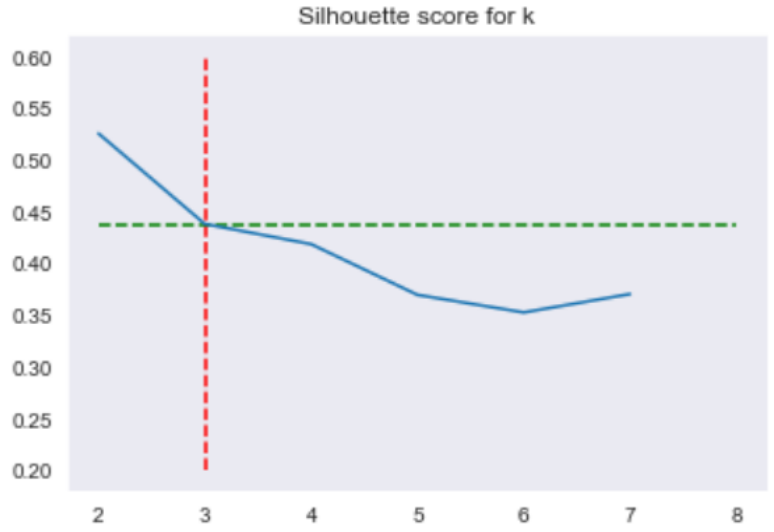
Kmeans clustering

- Create new dataframe with respect to new PCs.
- Based on elbow curve and silhouette score optimal clusters is 3.
- We will use standard sklearn package KMeans to cluster the data.
- Count of countries assigned to each cluster are as below

Cluster#	Number of Countries
0	65
1	42
2	24



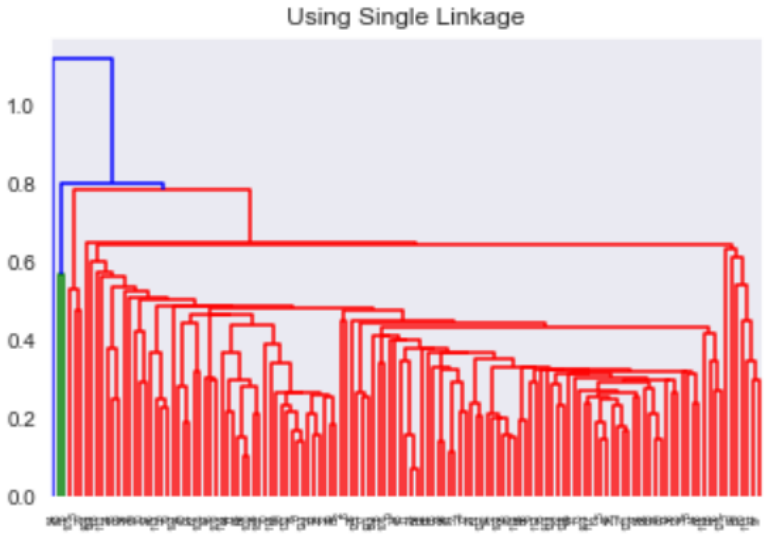


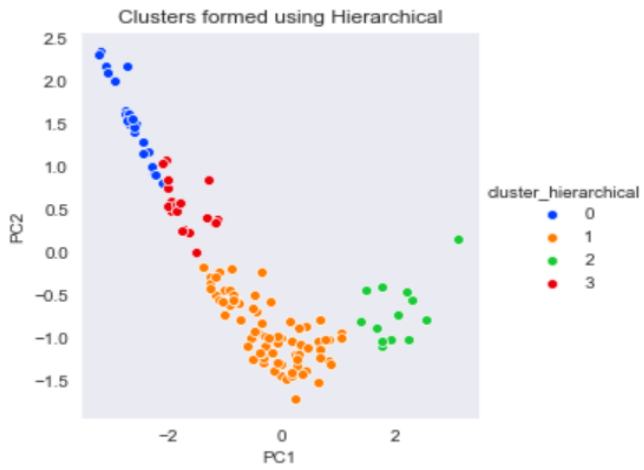


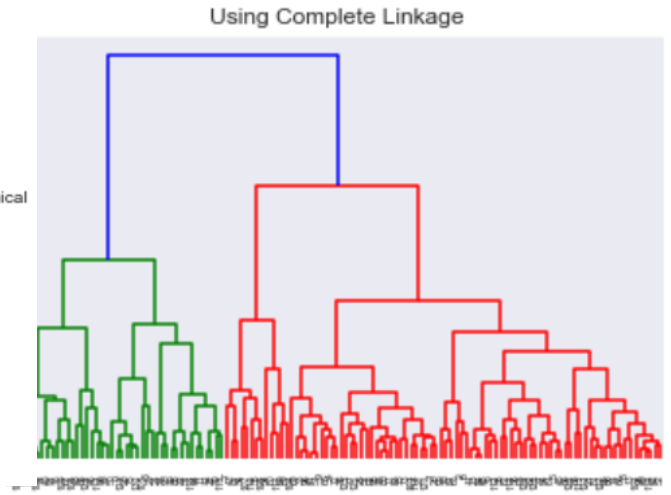
Clustering on PCA dataset (Hierarchical)

Hierarchical clustering

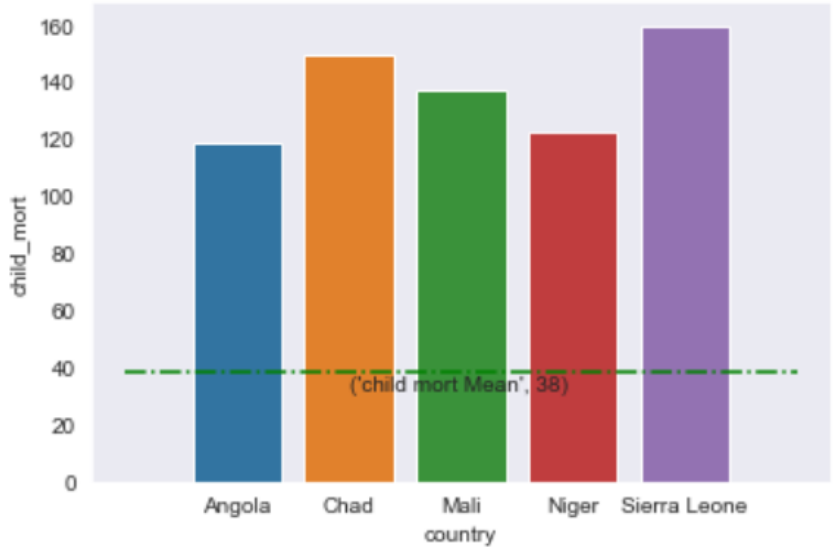
- Hierarchical clustering offers advantage as we do not have to pre-define number of clusters required.
- Using single linkage method we do not get well defined clusters. Single linkage uses shortest distance between points in 2 clusters.
- Using complete linkage methods gives us good clusters. Complete linkage uses maximum distance between points in 2 clusters.

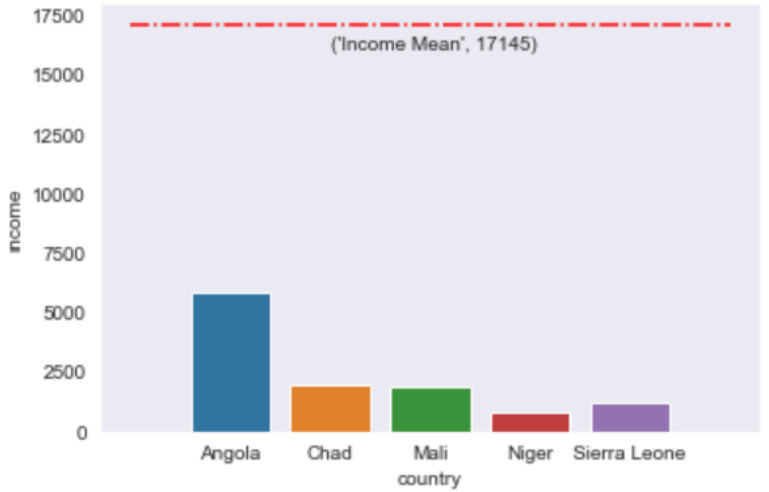


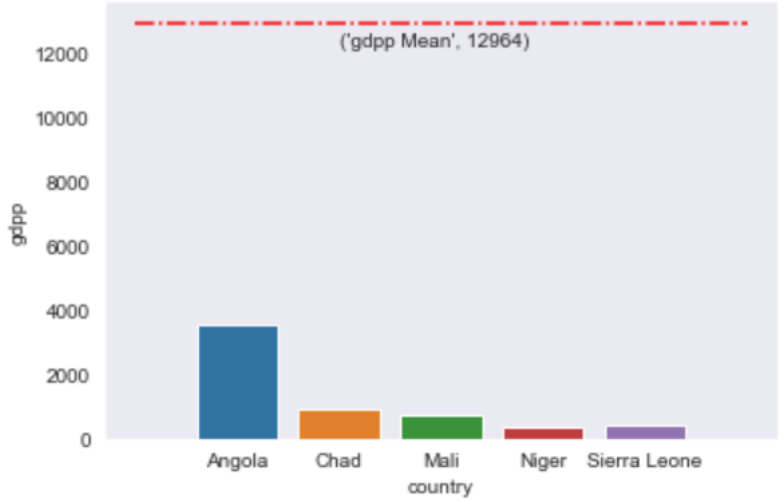




Conclusion







From Cluster 1 which has countries with lowest income, gdpp and highest child_mort, we will use child mortality and life expectancy factors to identify the countries for funds. These additional funds can be utilized for improving the health condition of the country.

Top 5 countries that need aid are

1. **Sierra Leone**
2. **Chad**
3. **Mali**
4. **Niger**
5. **Angola**