Clustering & PCA Assignment

GOAL :CATEGORIZE THE COUNTRIES USING SOCIO-ECONOMIC AND HEALTH FACTORS THAT DETERMINE OVERALL DEVELOPMENT OF A COUNTRY. AND THUS ULTIMATELY HELP NGO TO LIST DOWN THOSE COUNTRIES WHICH ARE IN DIREST NEED OF HELP.

Submitted By:
Vaishali Papneja

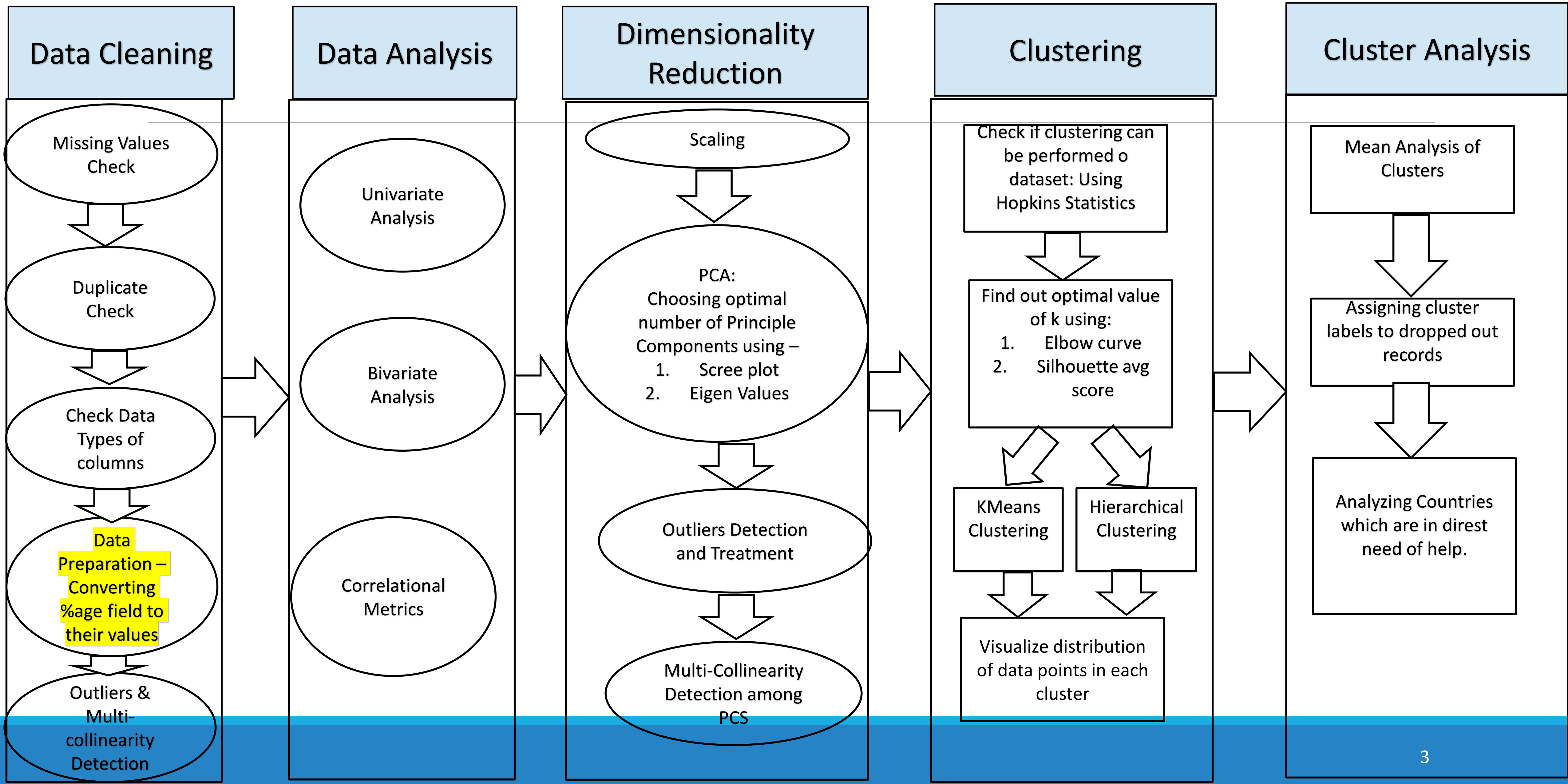
PROBLEM STATEMENT

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.

After the recent funding programs, 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.

We as Data Analyst need to suggest the countries which CEO needs to focus the most.

Steps Involved in Analysis



Data Cleaning

Missing Values Check

Duplicate Check

Check Data Types of columns

Data Preparation – Converting %age field to their values

Outliers & Multi-collinearity Detection

Data Analysis

Univariate Analysis

Bivariate Analysis

Correlational Metrics

Dimensionality Reduction

Scaling

PCA:
Choosing optimal number of Principle Components using –
1. Scree plot
2. Eigen Values

Outliers Detection and Treatment

Multi-Collinearity Detection among PCS

Clustering

Check if clustering can be performed o dataset: Using Hopkins Statistics

Find out optimal value of k using:
1. Elbow curve
2. Silhouette avg score

KMeans Clustering

Hierarchical Clustering

Visualize distribution of data points in each cluster

Cluster Analysis

Mean Analysis of Clusters

Assigning cluster labels to dropped out records

Analyzing Countries which are in direst need of help.

Data Analysis

1. After careful examination of Data – dictionary, I found that there are 3 columns – exports, imports and health which are in percentage form of Total GDPP.

As for doing PCA and cluster analysis later, we need to know exact values of these columns as it may be possible to have least value of GDP for a country but export percentage is highest and many such cases. In that case, our Principle components will not be able to capture actual information of such features.

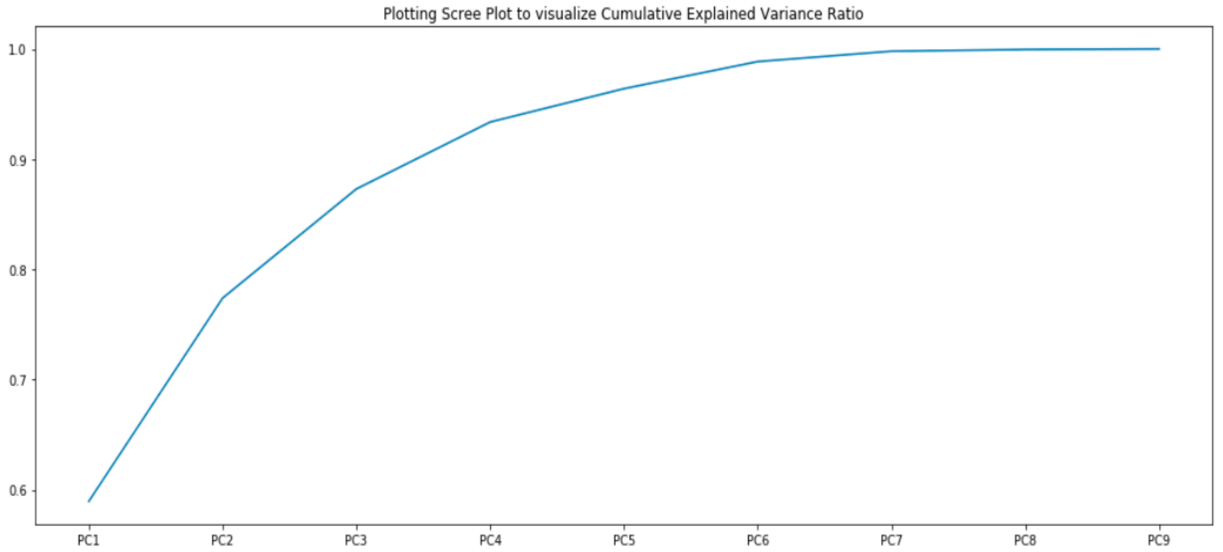
- Now, to convert these %ages to actual values, I have used GDPP feature as I have found on Google that GDPP is a method of measuring GDP of a country.

2. Outliers in our Dataset: After plotting various scatter plots between different feature variables, I realized that outliers present in our dataset are a case of either most poor condition of a country or most good condition of a country.

Results of PCA

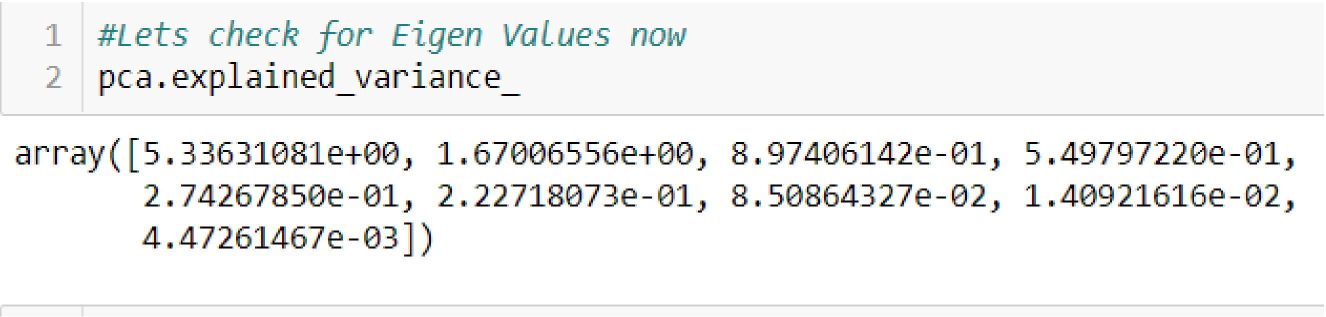
Selecting Principle Components which captures Maximum Information

Scree Plot



From Scree plot,

I see that almost 87% of the variance is explained by 3 principle components. But also, 4 PCs can explain 93% of variance. Now to decide on optimal number of PCs to be taken, I went ahead and checked eigen values



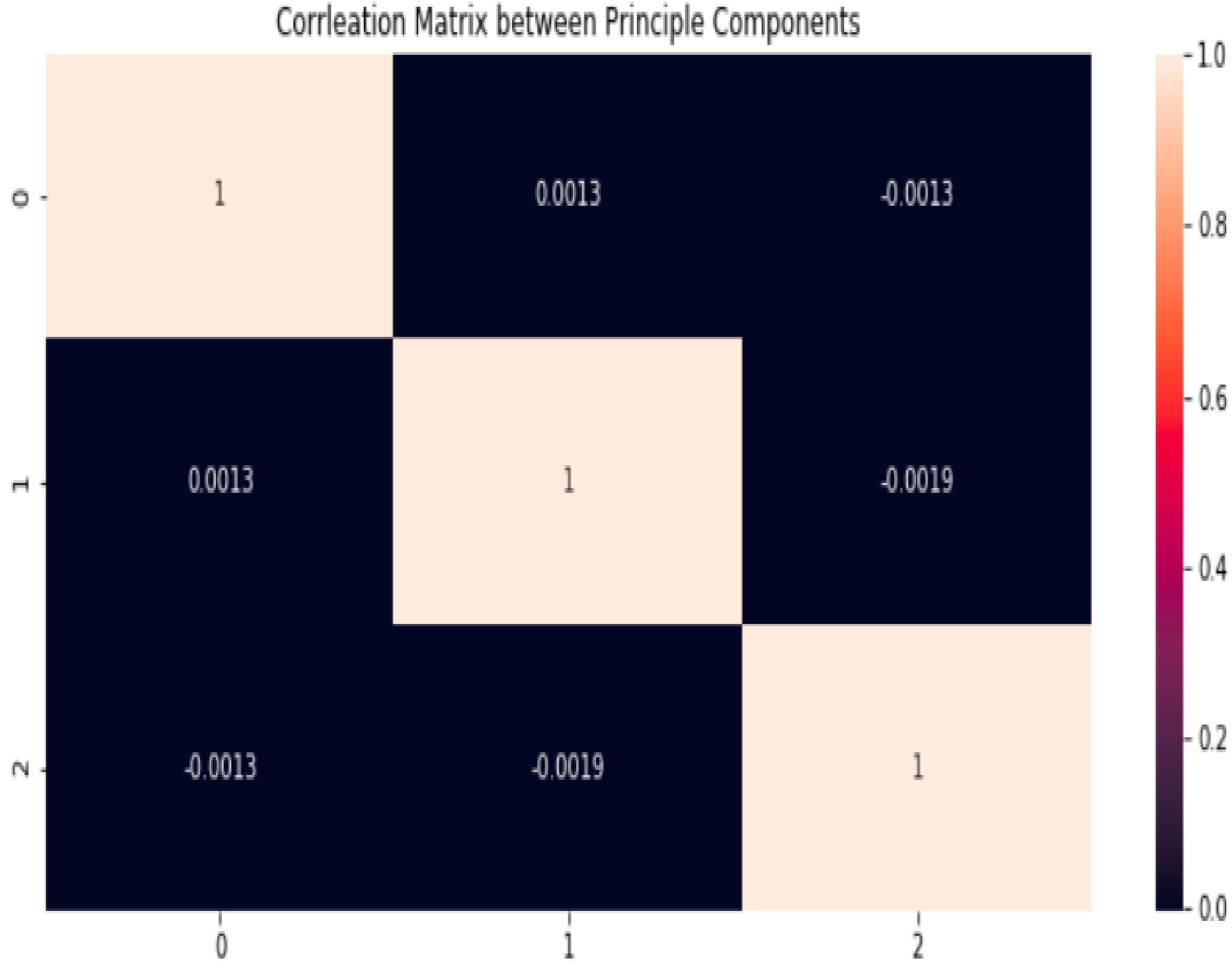
```
#Lets check for Eigen Values now
pca.explained_variance_
```

```
array([5.33631081e+00, 1.67006556e+00, 8.97406142e-01, 5.49797220e-01,
       2.74267850e-01, 2.22718073e-01, 8.50864327e-02, 1.40921616e-02,
       4.47261467e-03])
```

Through Eigen Values,

I see that PC1 and PC2 have eigen values greater than 1,and PC3 has eigen value almost equal to 1 but PC4 has Eigen value as 0.5 which is less than 1
which means this PC doesn't have much of a variation. So PC1, PC2 and PC3 are optimal principal components for our dataset which have most of the information.

Correlation between Principle Components

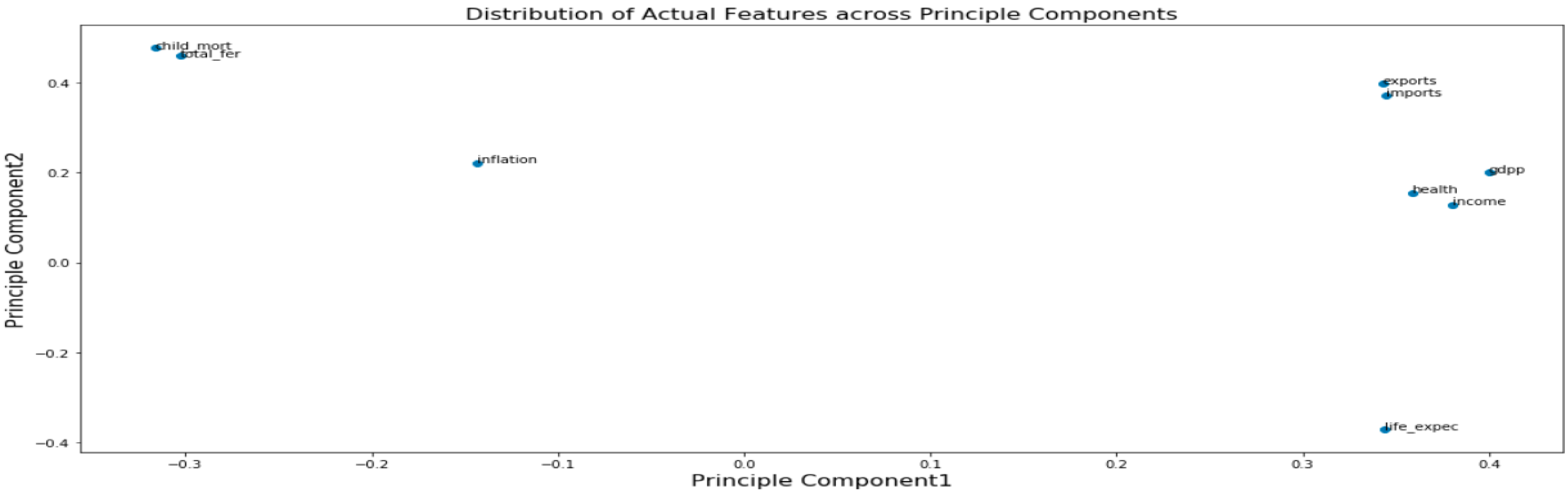


Corrleation Matrix between Principle Components

	0	1	2
0	1	0.0013	-0.0013
1	0.0013	1	-0.0019
2	-0.0013	-0.0019	1

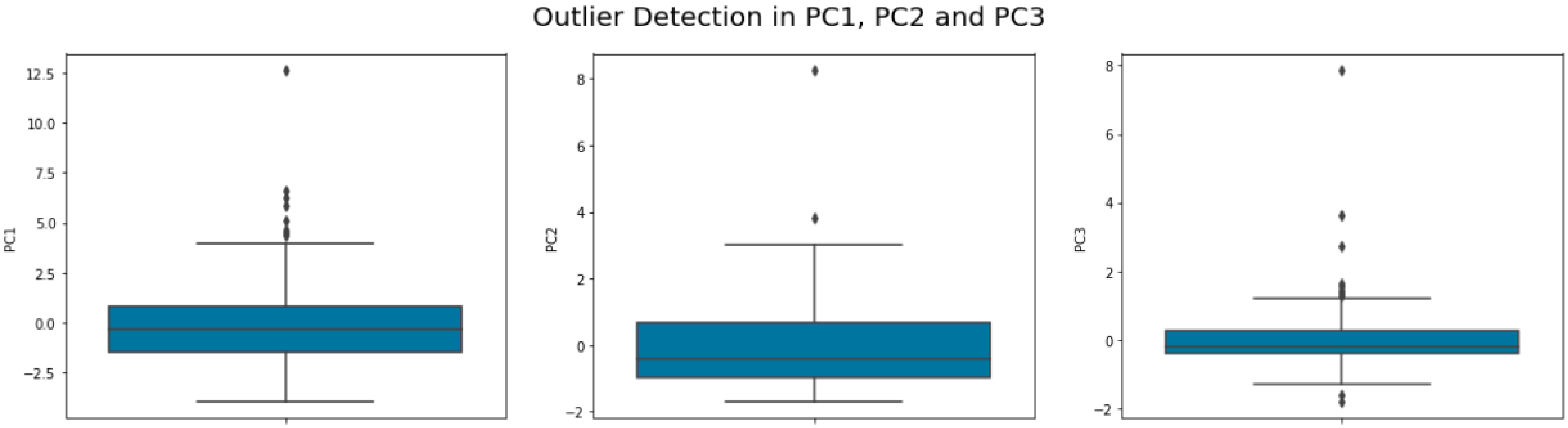
I see that there is minimum or say, no correlation between my Principle Components. This is because, PCs are Orthogonal to each other.

Distribution of Actual Features Across PCs



I see that all of the feature variables are nicely explained by principle components - Child_mort and total_fer are explained by PC2 and income, GDPP, health, life_expec, imports, exports are explained by PC1.
And inflation is explained by combination of PC1 and PC2.

Outliers Detection



Above plots clearly indicates presence of outliers. As previously explained, these outliers can be a result of extreme poor condition of a country or extreme good condition of a country.

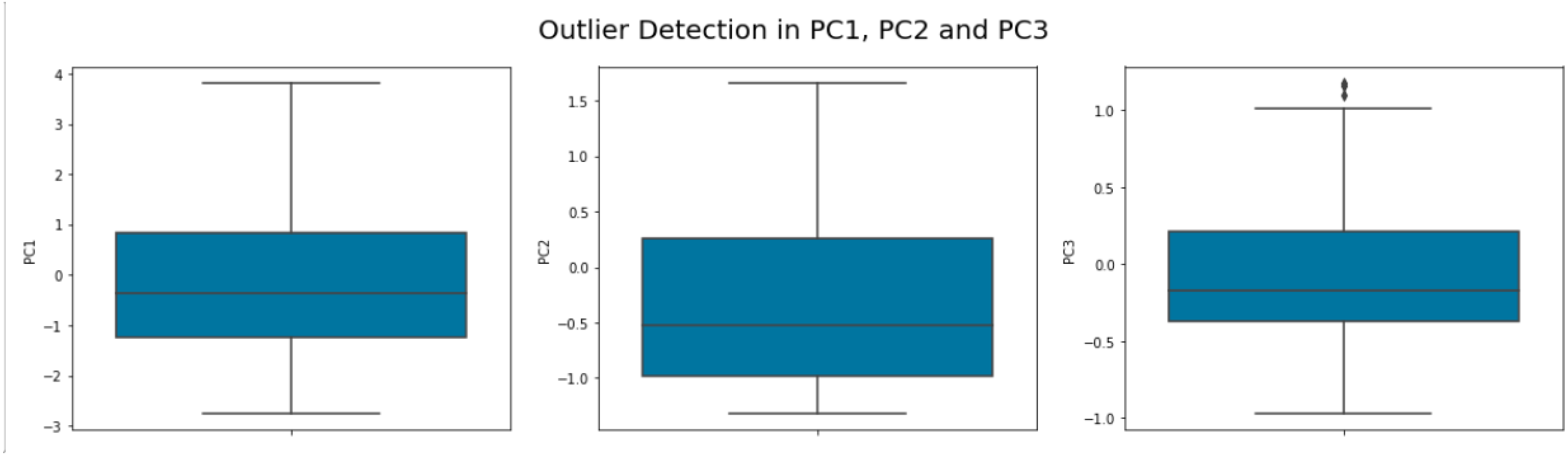
But to proceed with Clustering, we need to remove these outliers and later we can analyze dropped records on the basis of Mean, Min and Max value analysis of socio-economic factors of all clusters.

Outliers Treatment

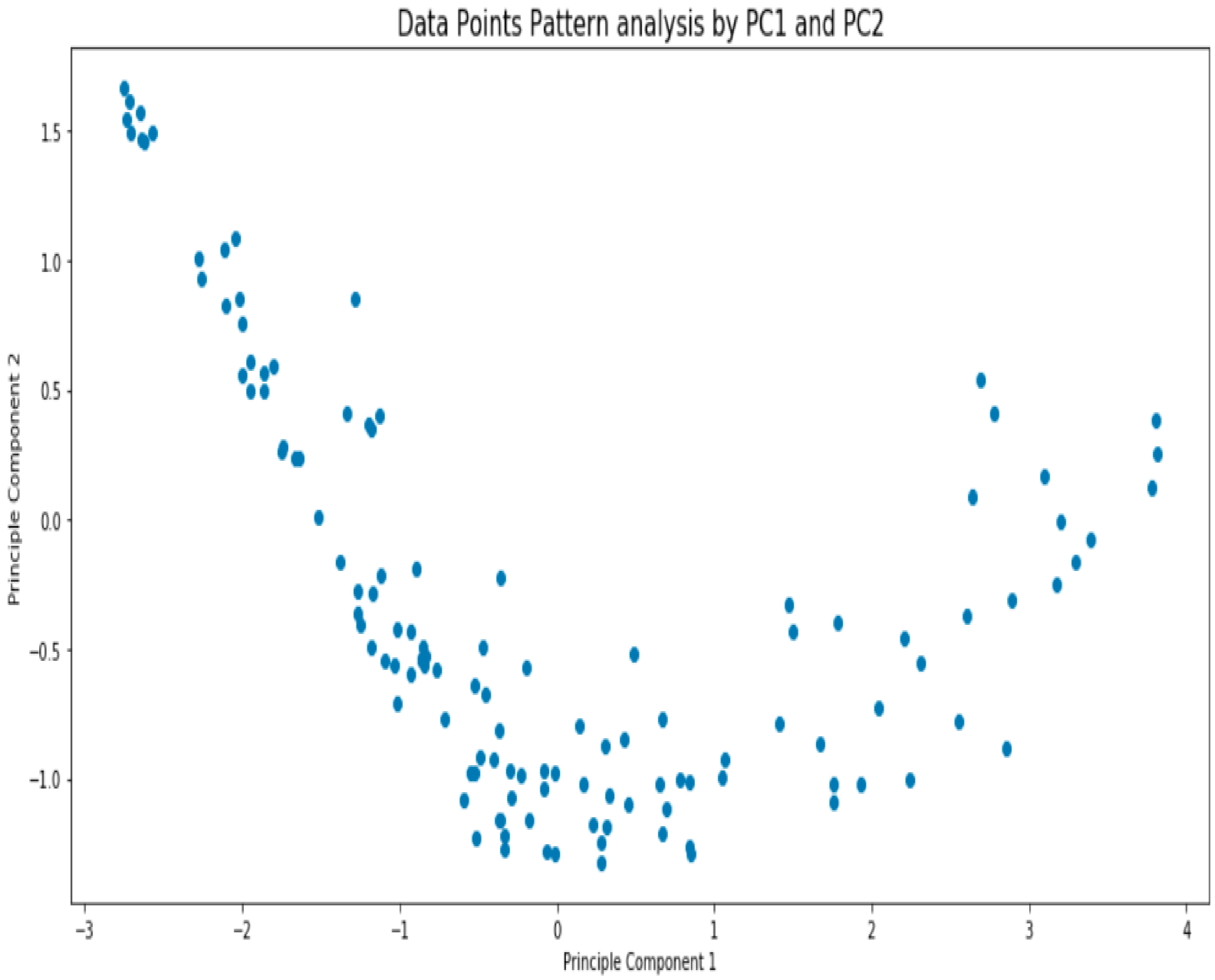
To Treat outliers present in our dataset, I have considered only those data points which lie in between 5% and 95% data on the basis of:

1. Child_mort
2. Income
3. GDPP

After doing so, I am left with 125 countries and there are no outliers in the dataset.



Visualizing Data Points on the basis of PC1 and PC2



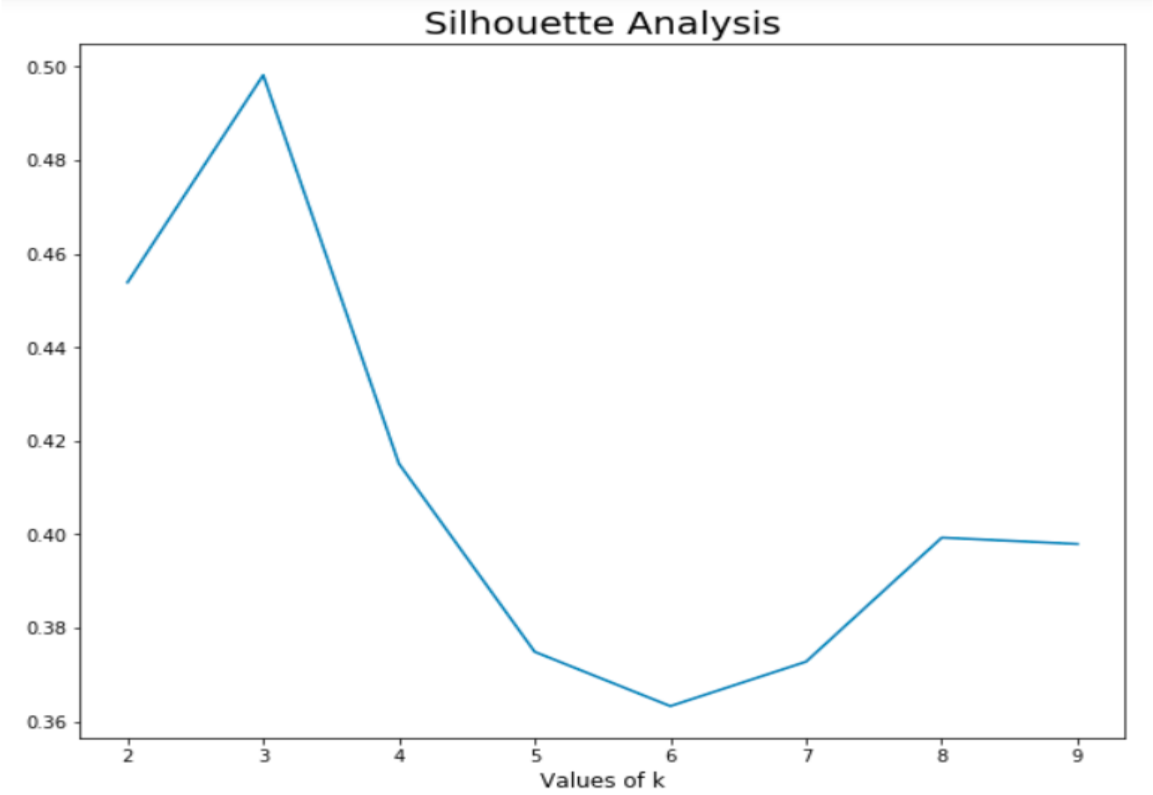
Here I can clearly see some nice cluster formation.

Results of Clustering

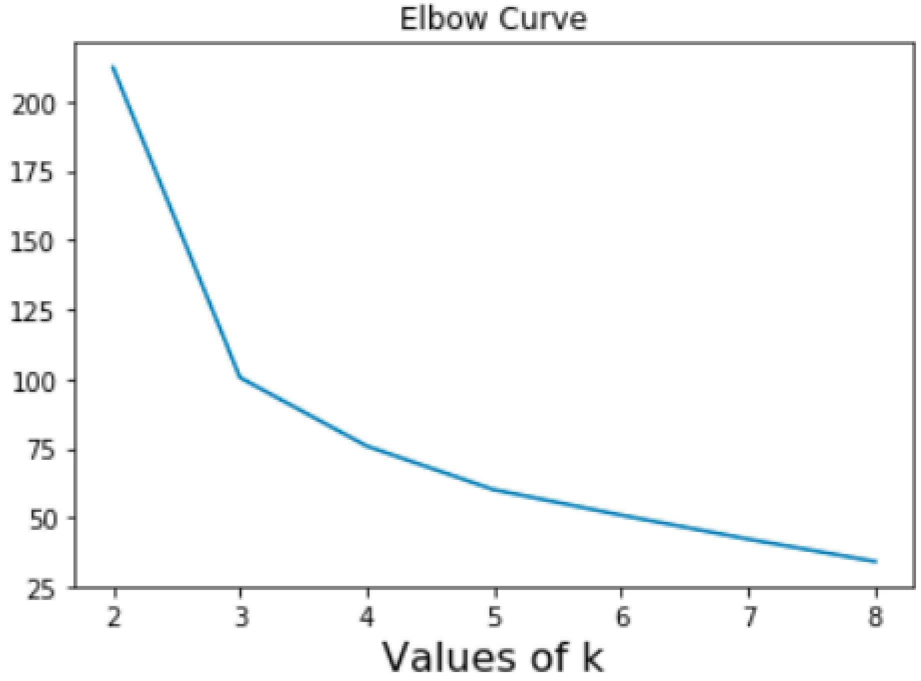
Check if dataset is suitable for Clustering - Hopkins Statics

Before proceeding with Clustering, I first verified if dataset is suitable for clustering. For this, I checked Hopkins score and found that score for our data set is 0.82 which shows that dataset is good enough to proceed with clustering.

Choosing Optimal value of k



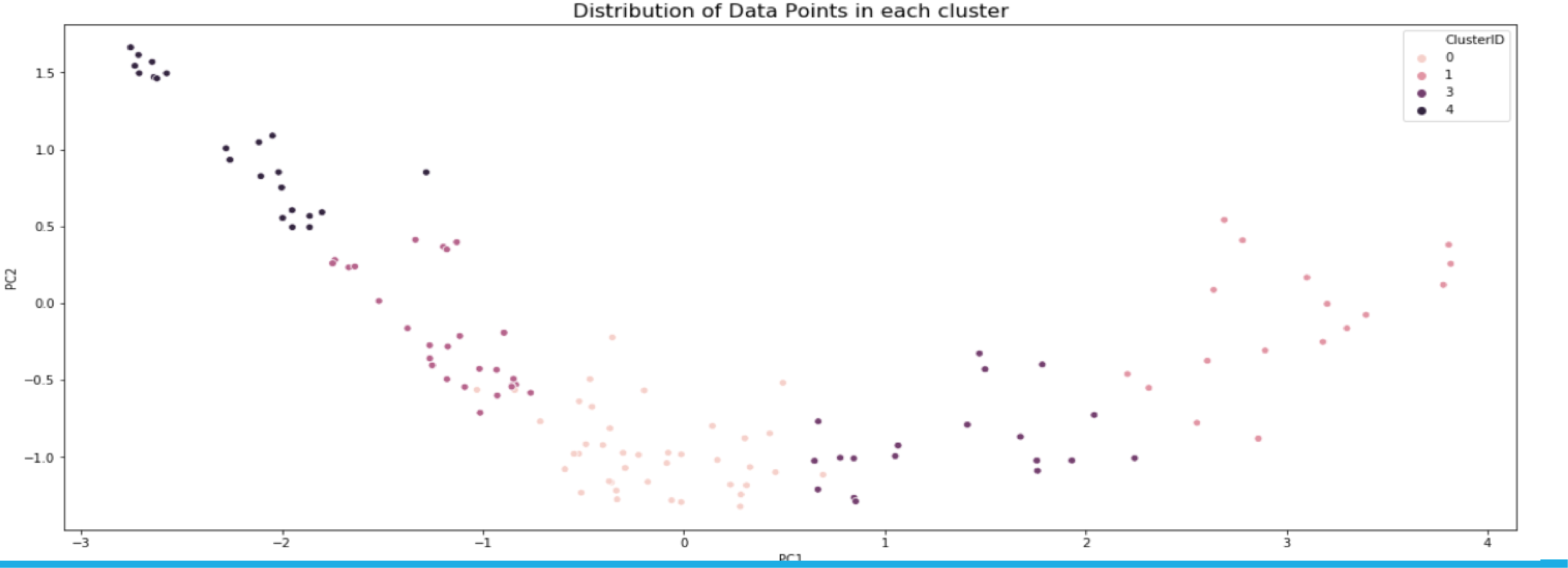
Silhouette score tells us the average distance between data points across different clusters.
So, we always choose maximum silhouette score so that data points of different clusters behave differently.



Elbow curve tells us distance of data points which lie in a cluster.
We always choose value of k such that distance between data points of a cluster is minimum.
By above plot, I can say that k=3 or 5 is optimal value for number of clusters.

KMeans Clustering: Distribution of Data Points in Each Cluster

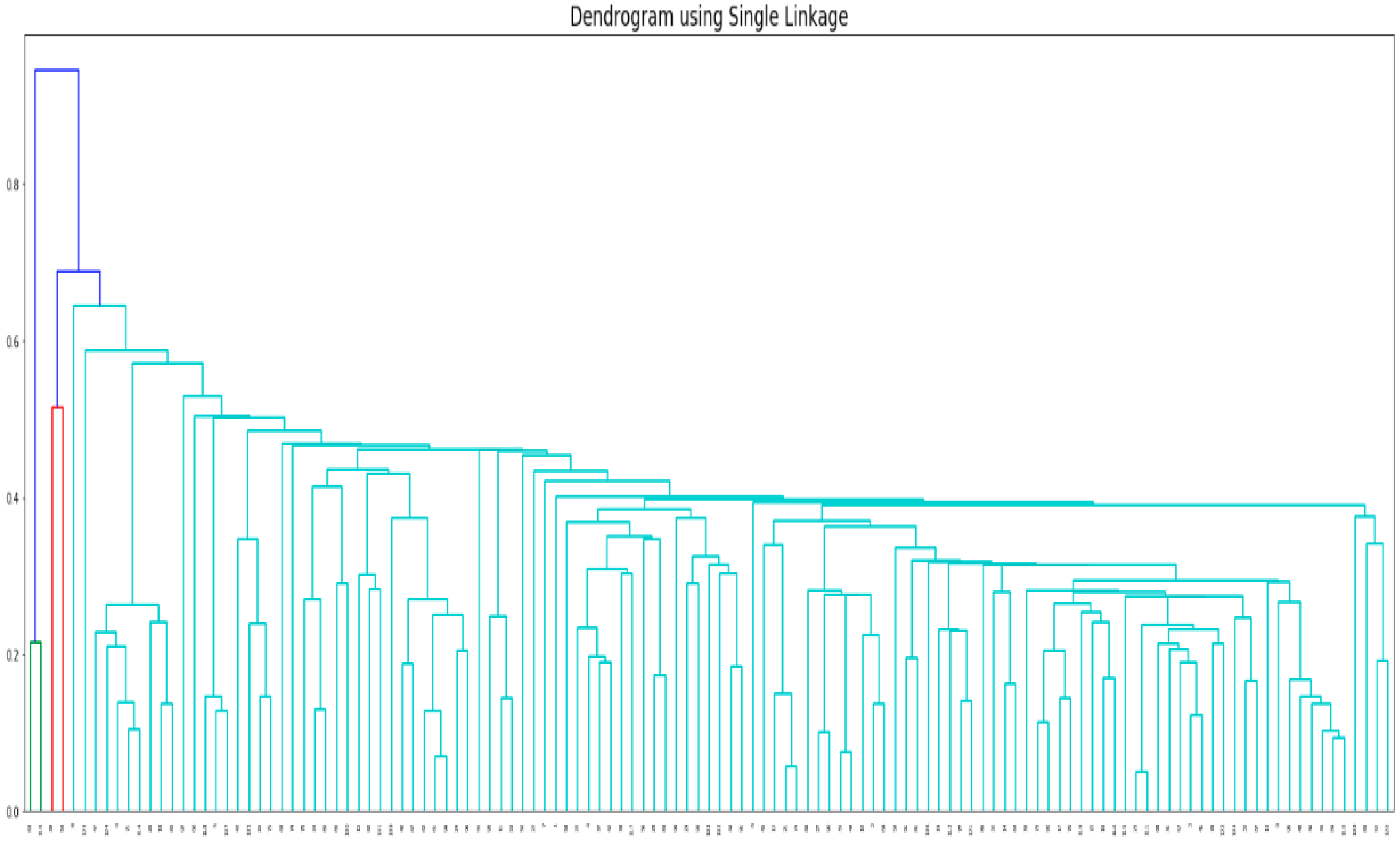
I created clusters by using k=3 and k=5. There I realized that K=5 is able to give me more optimal results. As with k=3, I got 1 clusters with huge variance in mean value in comparison of other clusters.
Also, as we have limited money for investment, we want to first get those countries which are in direst need to help, so distributing these countries into more clusters looks beneficial.

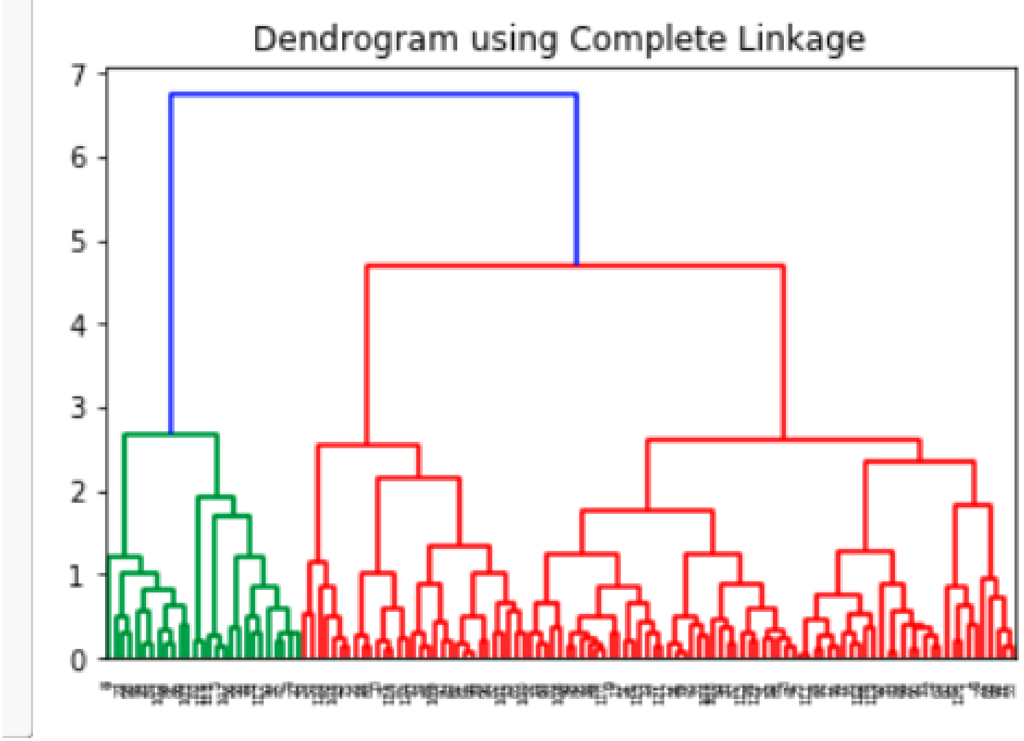


Hierarchical Clustering

Here I have plotted Dendrograms using Single as wells as complete linkage.
Single Linkage – shows us least distance between data points of 2 clusters whereas
Complete Linkage – shows us the farthest distance between data points of 2 clusters.

Here also, I found that k=3 is an optimal value for number of clusters.





Cluster Analysis for Outliered Clusters

To assign clusters to our dropped off countries, I checked mean values of different features in each cluster and came up with below logic to assign them to a cluster:

On the basis of child_mort:

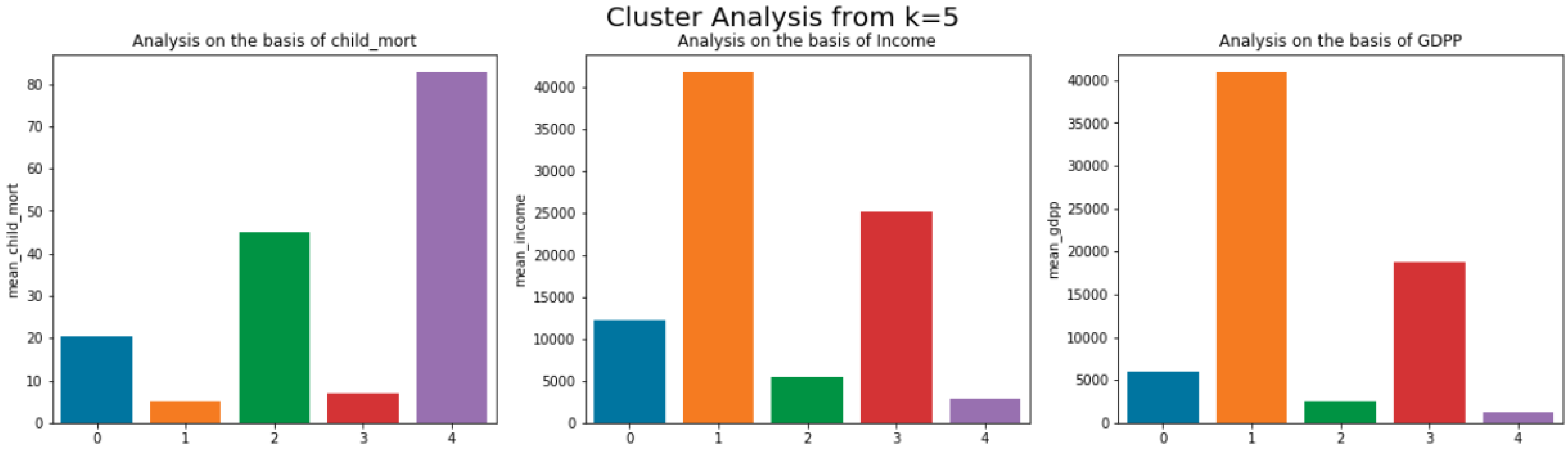
if(child_mort >= 55.2): then assign cluster as 4
else if(child_mort value >=17.4): assign cluster as 2
else if(child_mort value >=5.5): assign cluster as 0
else if(child_mort value >=3.2): assign cluster as 3
else if(child_mort value >=2.6): assign cluster as 1

After assigning cluster labels on the basis of child_mort ratio, I cross-verified the labels based on Income criteria which is as below:

1. if income >= 28300: then cluster 1
2. if income >= 15300: then cluster 3
3. if income >= 3910: then cluster 0
4. if income >= 1350; then cluster 2
5. if income >=700, then cluster 4

Upon verification, I got the same results as after assigning clusters on the basis of child_mort.

Mean Analysis of Clusters



From above plots, it is clear that countries in cluster 4 are in dire need of help. As they have

1. Most child_mort ratio
2. Lease income
3. Least GDPP

Analyzing Countries which are in direst need of Help

By now, I have found a cluster of countries which are in direst need of help. Now, there are almost 30 countries present in that cluster. But we are looking for topmost countries which require immediate help.
To do so, I have calculated mean of child_mort, income and gdpp and chosen only those countries where :

Criteria:

Child_mort rate is > child_mort_mean(93.6) and
Income < income_mean (1820)
OR GDPP < gdpp_mean (2365)

Here I have applied ‘or’ condition between income and GDPP because after taking a close look at data set, I found that there are countries where Income is little more than mean but GDPP is very less.
As countries health can be determined by GDPP, I have taken all such countries into account.

Final List of Countries which are in direst need of Help

As we have a limited budget, so want to know Top 10 countries first which needs help urgently.

To find the same, I have sorted the list of countries on the basis of Child_mort in Descending order and GDPP in Ascending order, I could find TOP 10 countries which are in urgent need of help.

Below is the list:

1. Haiti
2. Sierra Leone
3. Chad
4. Central African Republic
5. Mali
6. Nigeria
7. Niger
8. Congo, Dem. Rep.
9. Burkina Faso
10. Guinea-Bissau

If HELP NGO wants to invest $10M, they should focus on above countries first.