PCA ASSIGNMENT

BY

Kvs Hemendra Sai

PROBLEM STATEMENT & ANALYSIS

- HELP International, an NGO is has just raised 10 Mn USD and is tasked with allocating for development of underdeveloped countries.
- Since they will have to choose between countries to allocate funds appropriately, they were provided with data of 167 countries with various attributes as variables such as GDP per capita, Child Mortality, Income, Inflation etc.
- Our task is to identify a cluster of countries which need funding based on the variables from the datasets.
- Countries which require funding are usually the ones which are doing poorly on the economic indicators. They are usually associated with low GDP per capita, High Child Mortality, Low income, Low inflation, Lower life expectancy etc.
- Based on the above analysis, we need to find a cluster of countries which are doing poorly on the economic front.

PROCESS & APPROACH

DATA PREPARATION:

- Since there are no missing values in the data. We can proceed directly for Outlier treatment.
- It is observed that all the 8 variables have outliers and they are treated by considering the values above 99%'le as values at the 99th %'le and the values below 1%'le as values at the 1st %'le.
- After the Outliers are treated, all the values are standardized using standard scaler to perform PCA.

RESULTS OF PCA:

- After performing PCA, from the explained variance ratio, it is observed that more than 95% of the variance in the dataset is contributed by 5 principal components and hence dataset is transformed with 5 principal components to per form Kmeans clustering.

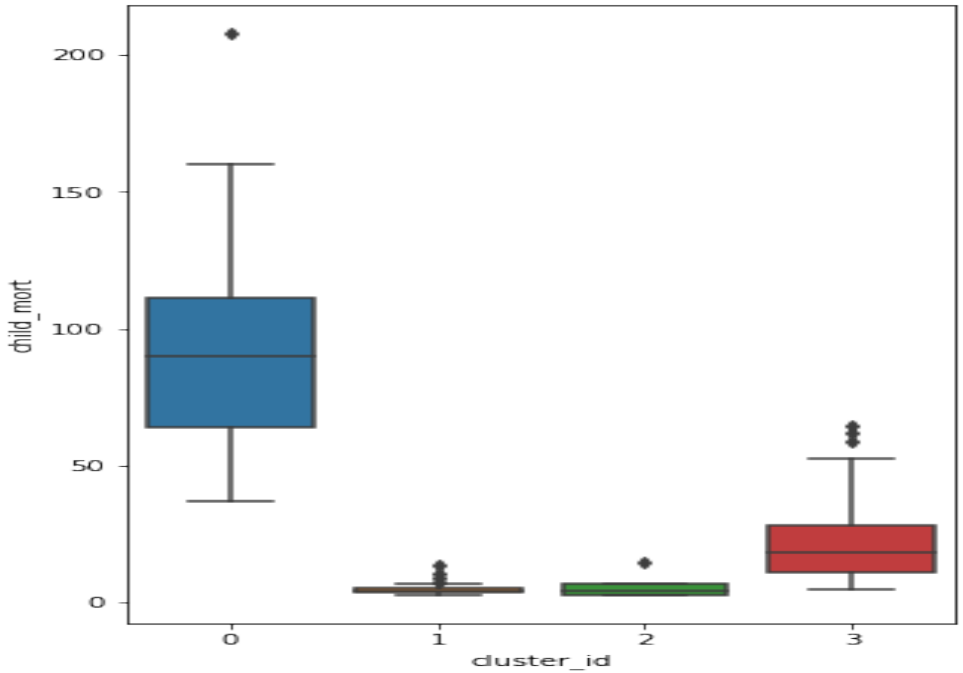
Kmeans Clustering:

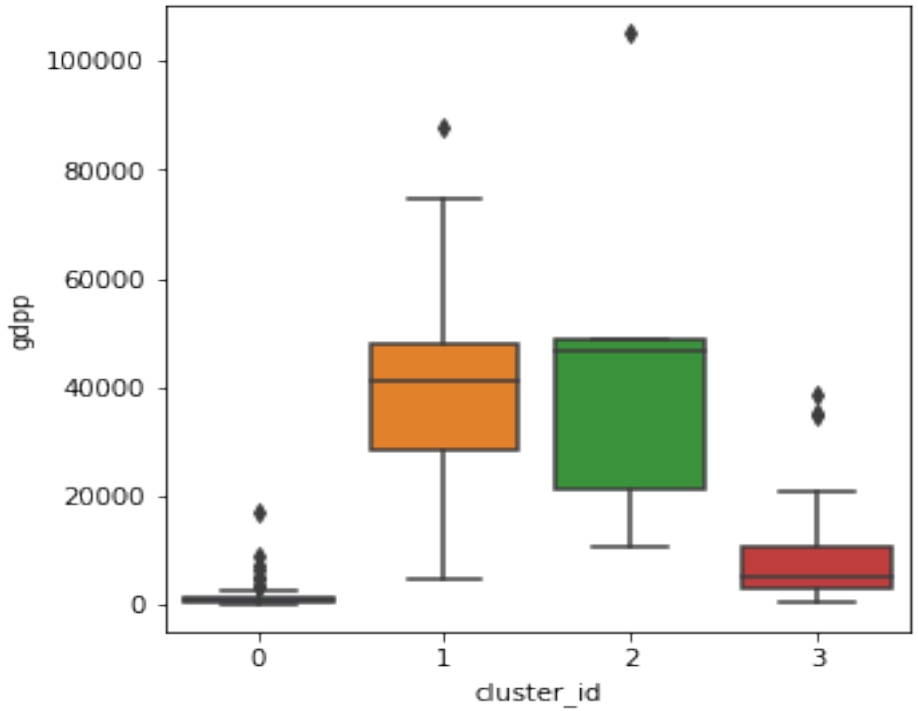
- Number of clusters in Kmeans clustering bis selected by plotting an elbow curve using sum of squared differences.
- From the curve, it is observed that SSD is drastically reducing as we move from 4 to 5 clusters indicating that the optimal number of clusters is 4.
- Also, Silhoutte number is calculated and it shows a good number above 0 for k = 4 with the silhouette number decreasing as we increase the number of cluster
- Therefore, Kmeans clustering is performed on the dataset with 5 principal components to obtain the clusters.

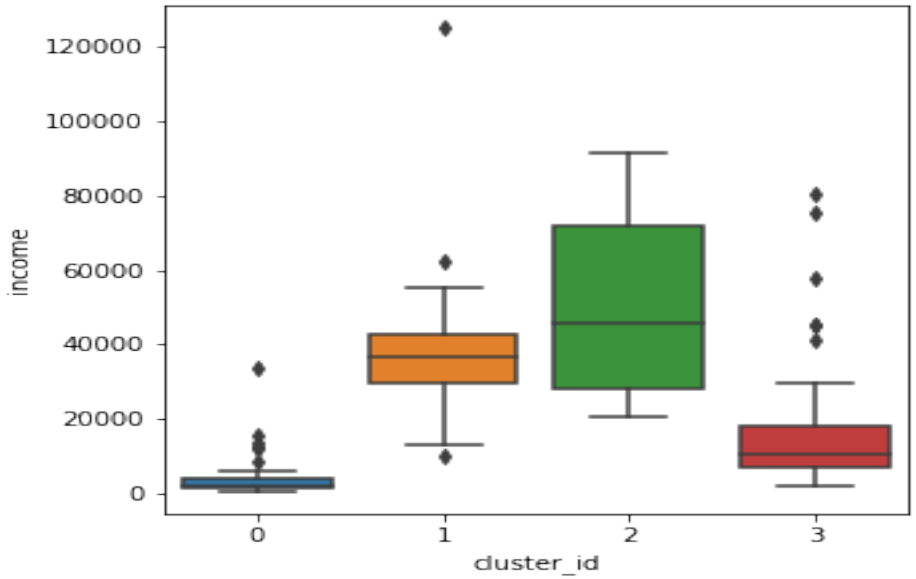
Results of Kmeans Clustering:

- Clusters thus performed with Kmeans clustering are analysed with 3 of the original variables i.e GDP percapita, Child Mortality & income.

- Following are the results of the analysis.







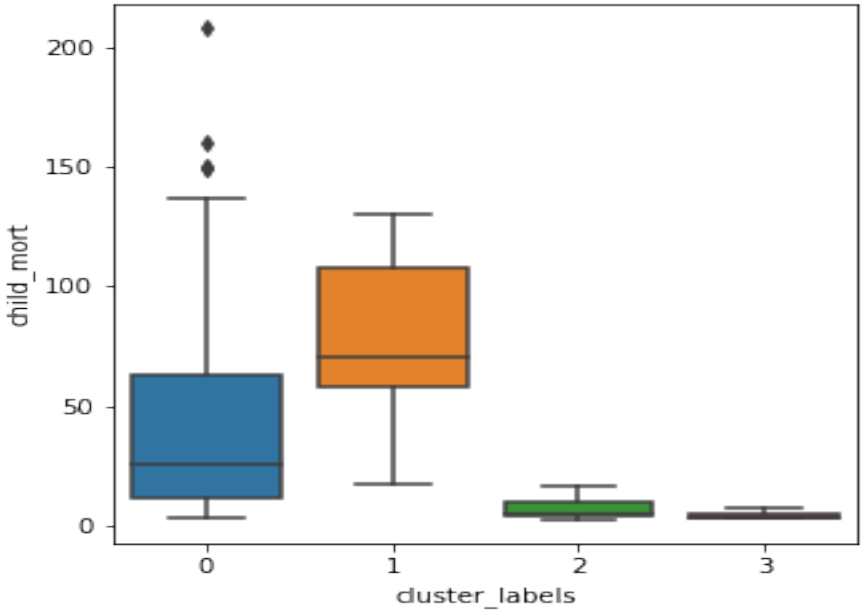
- From the results of Kmeans clustering, It is clear that countries in the cluster 0 exhibit the charecteristics of low income, low GDP per capita and high child mortality.
- Therefore, as per the Kmeans clustering it is advisable to allocate funds for the development of countries in cluster 0.

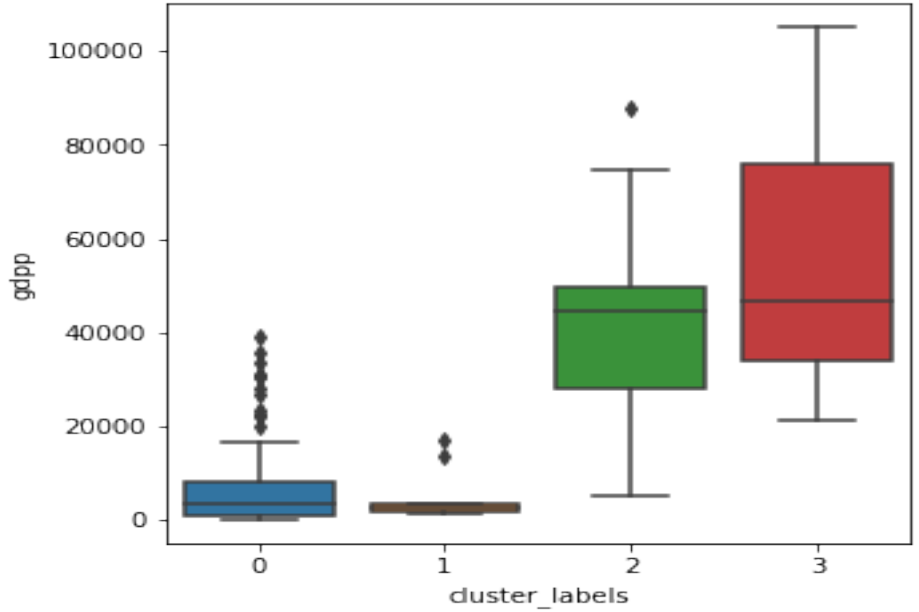
HEIRARCHIAL CLUSTERING:

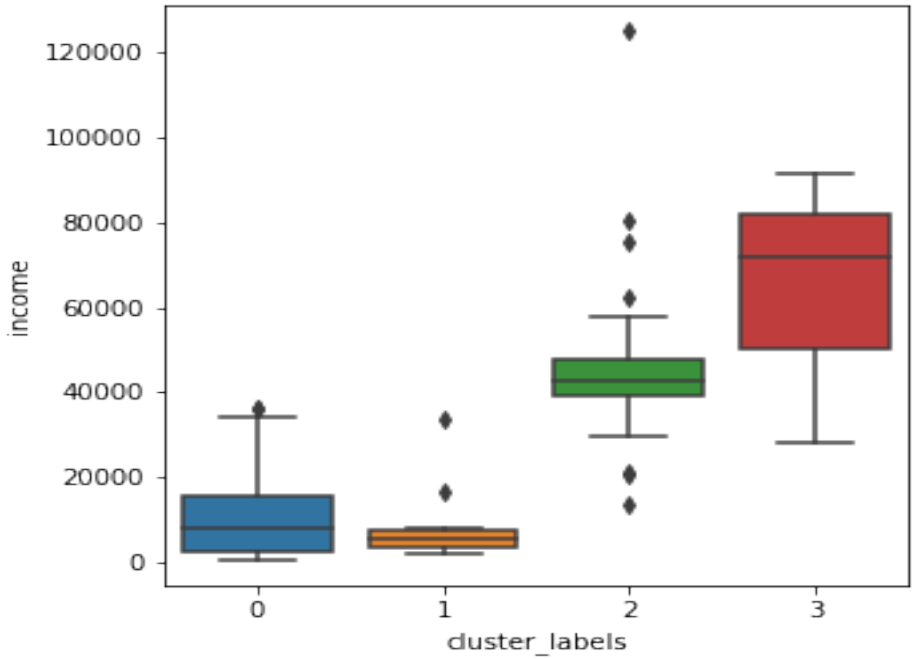
- For Heirarchial clustering, a Dendrogram is created using the complete linkage method. A Dendrogram is also created using a single linkage method but due to the limitations of single linkage method, clusters formed are not very clear.
- Dendrogram is cut in such a way that the number of clusters is 4.

RESULTS OF HEIRARCHIAL CLUSTERING:

- After performing Heirarchial clustering, clusters thus formed have been analyzed with respect to principal components and the three chosen original variables.





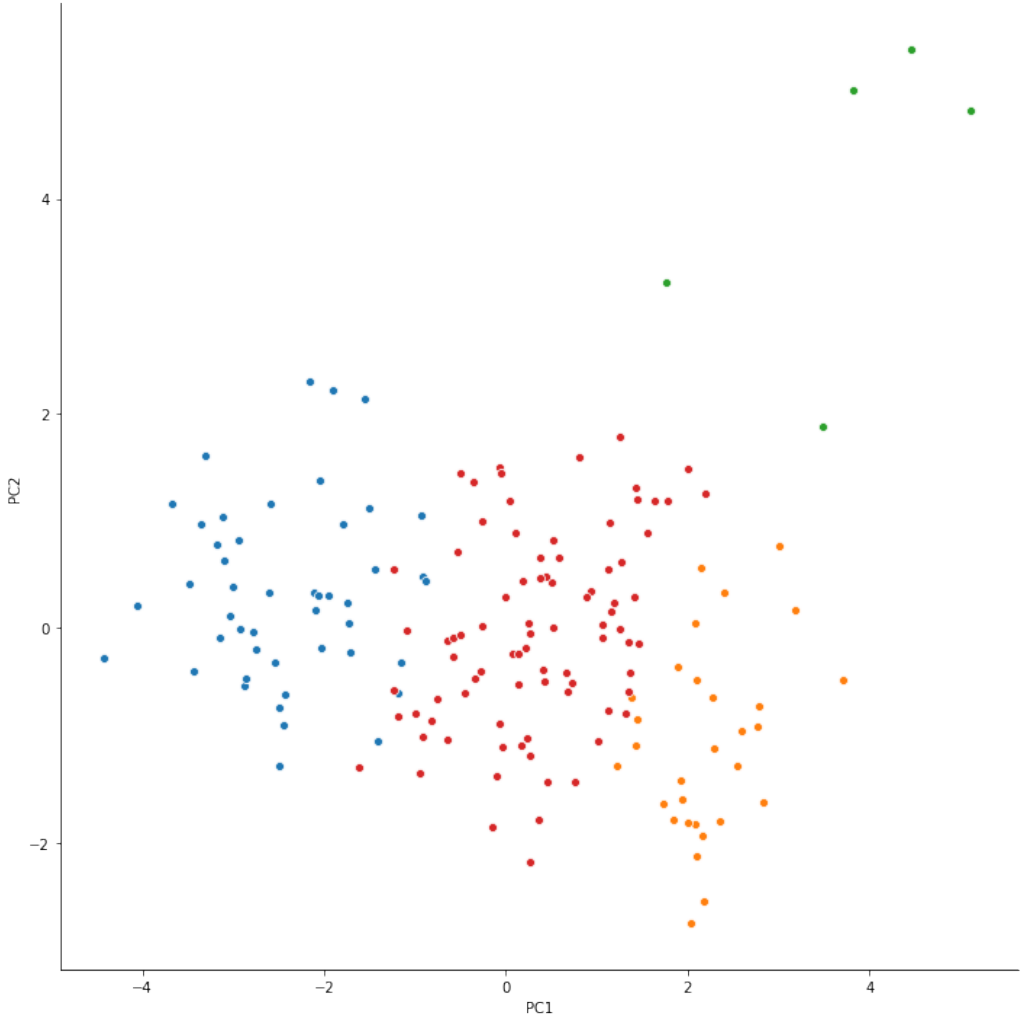


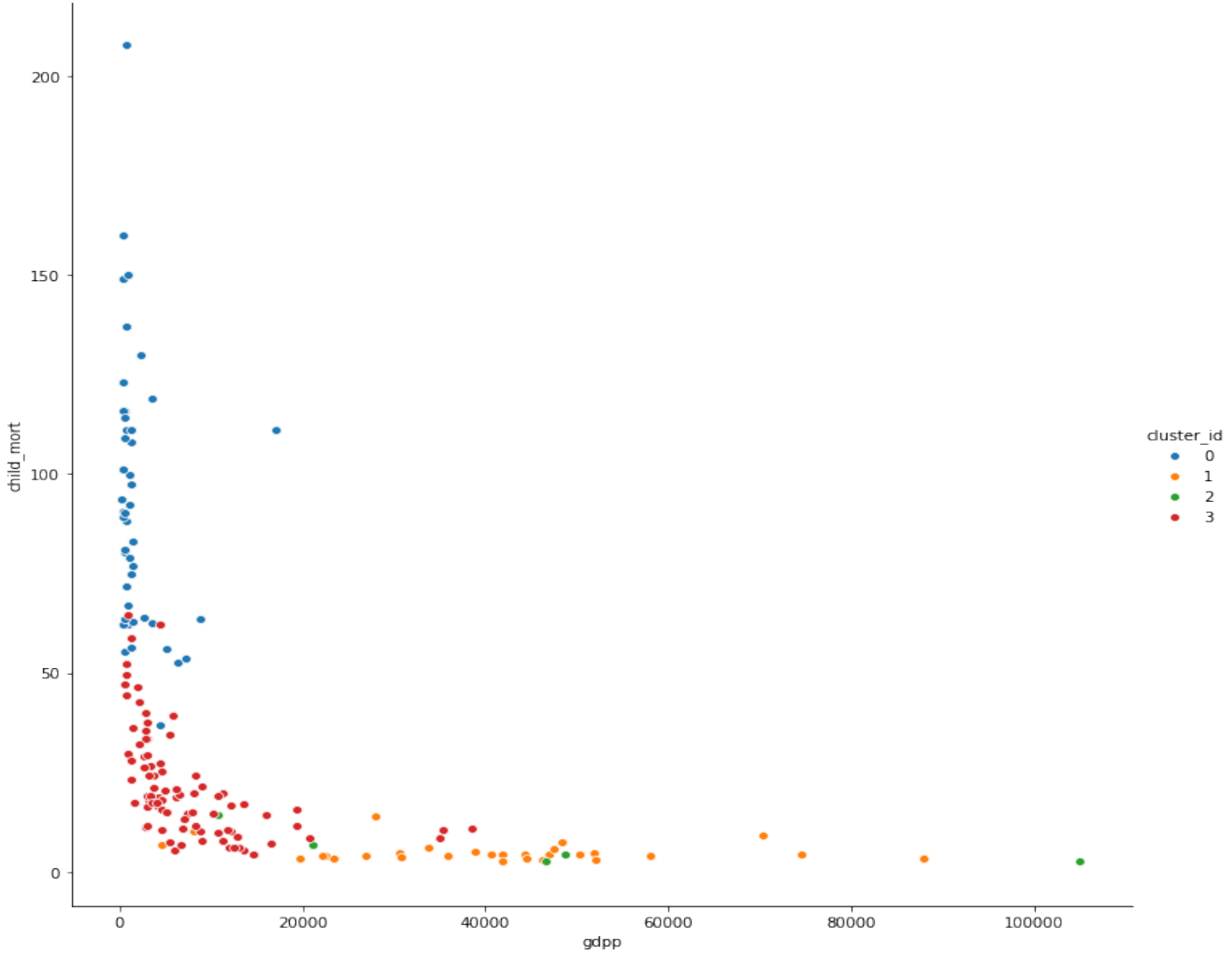
- From the above figures, we can observe that cluster number 1 has countries which have low GDP per capita, High child mortality and Low incomes.
- Therefore, as per Heirarchial clustering countries from cluster 1 are to be funded for development.

VISUALISATION OF CLUSTERS:

Kmeans Clustering:

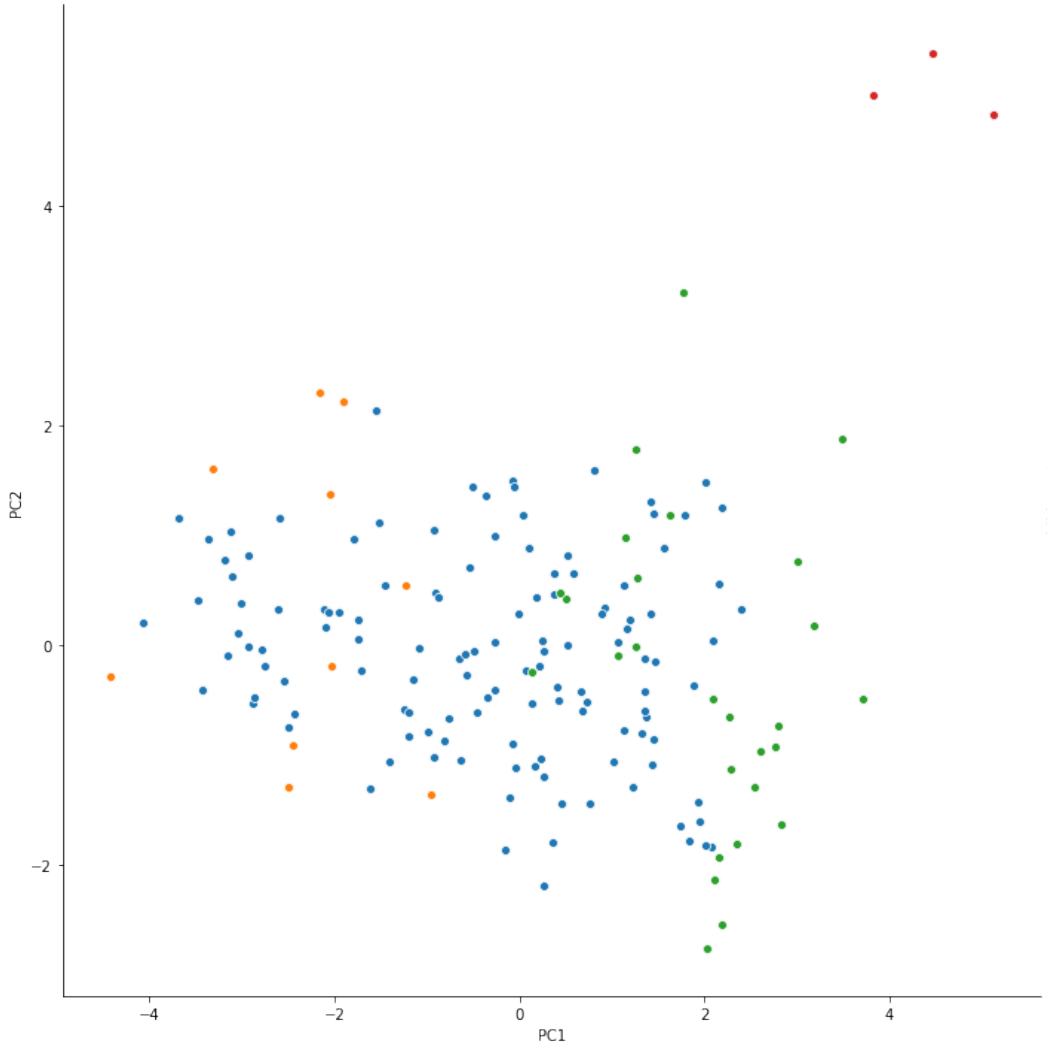
Visualisations of the clusters are performed with PC1,PC2 & Child_mort & gdpp

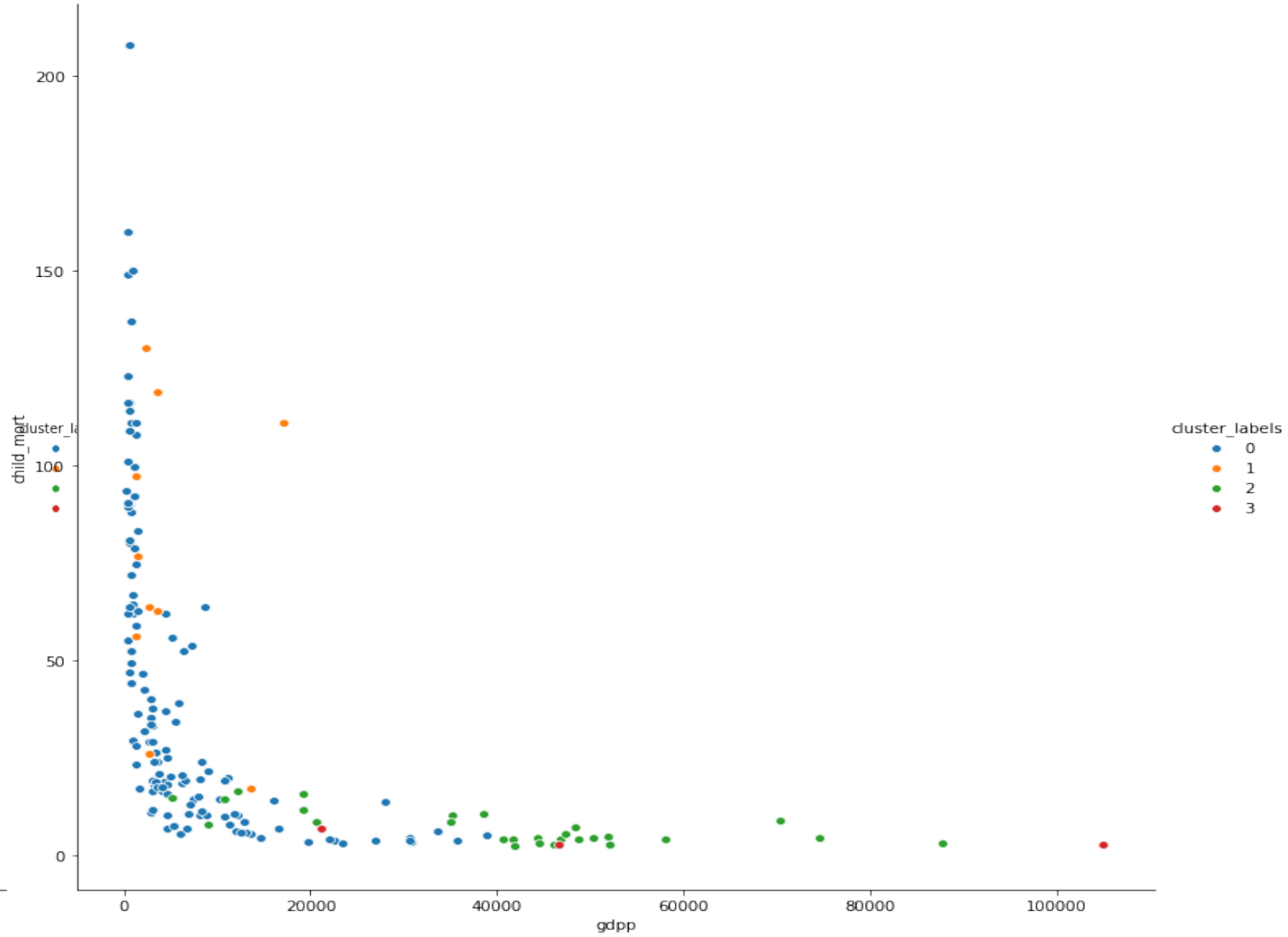




HEIRARCHIAL CLUSTERING:

Visualisations of the clusters are performed with PC1,PC2 & Child_mort & gdpp





Choosing the countries:

- Finally, as per the analysis, 47 countries require funding as per Kmeans clustering and 10 countries require funding as per Hierarchical clustering.
- It is also observed that 8 countries in the Hierarchical clustering also are present in the list of countries from Kmeans countries incdicating that these countries are in the direst need of funds.
- Therefore, the final list of countries are common in the Kmeans clustering and Hierarchical clustering. Final 8 countries are as follows.

1. Angola
2. Congo Rep
3. Equitorial Guinea
4. Mauritania
5. Nigeria
6. Sudan
7. Timor - Leste
8. Yemen