Clustering & PCA Assignment

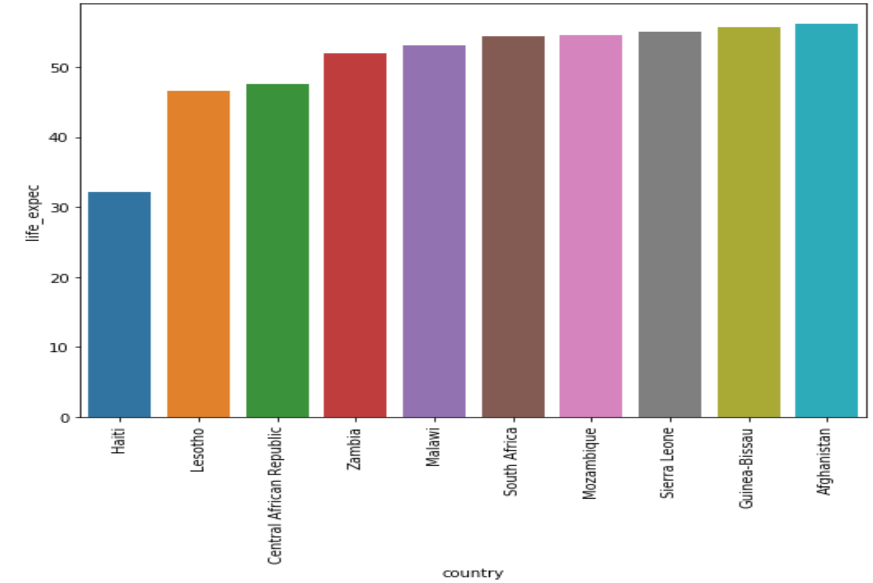
Karan Joseph

Problem Statement

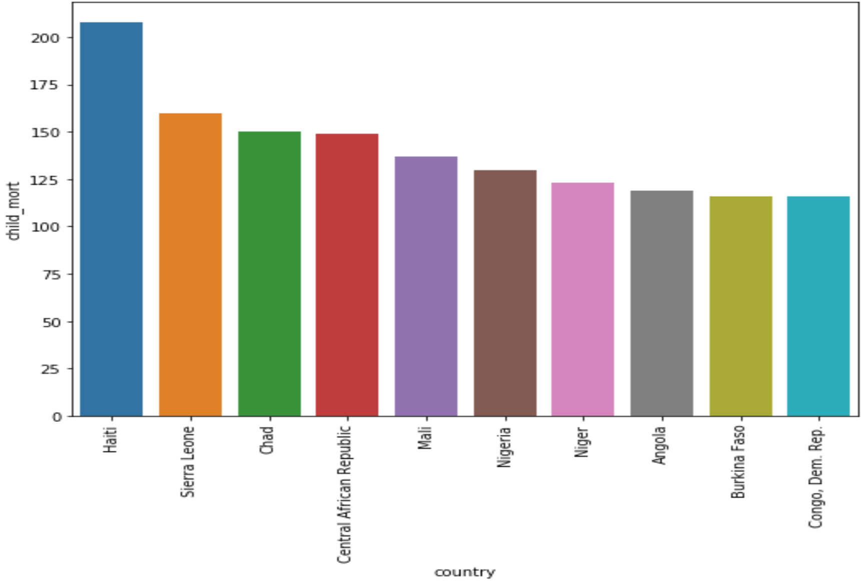
- HELP International (an NGO) has raised around $10 million and the CEO wants to know how he can effectively and strategically use this money to help the countries that are in the direst of needs.
- My objective is to successfully categorize the countries using some socio-economic and health factors such as gdpp, income, child-mortality rate etc. to determine the overall development of the country and then suggest my list of countries in need of immediate assistance.

Exploratory Analysis

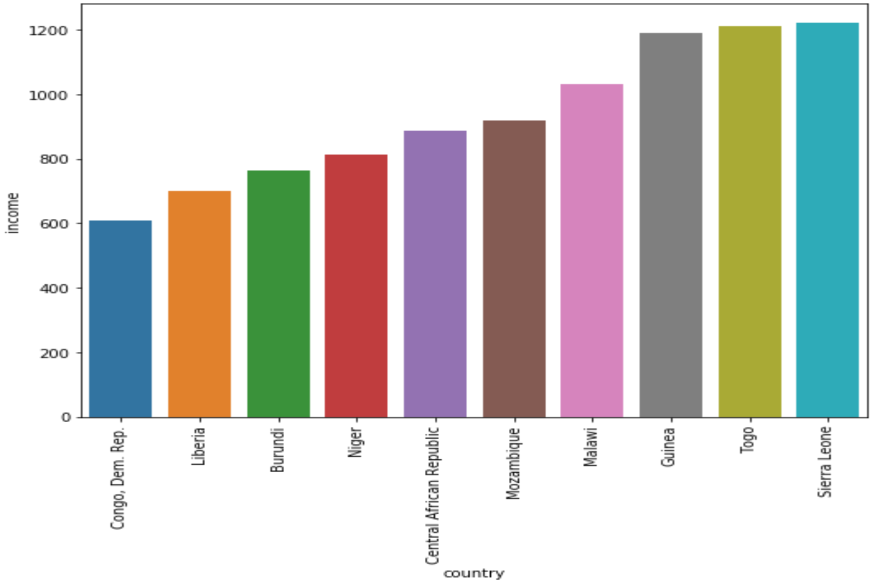
1.Countries with low health index



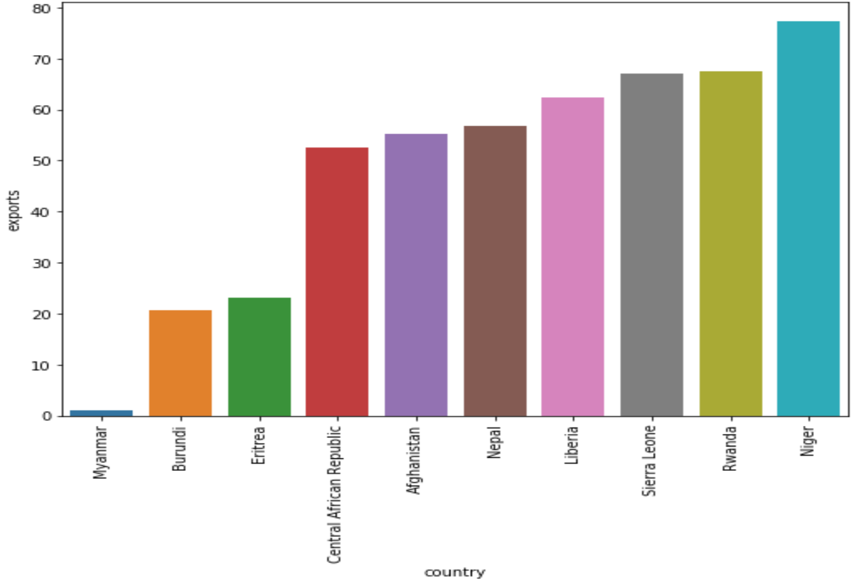
2.Countries with high infant mortality rate



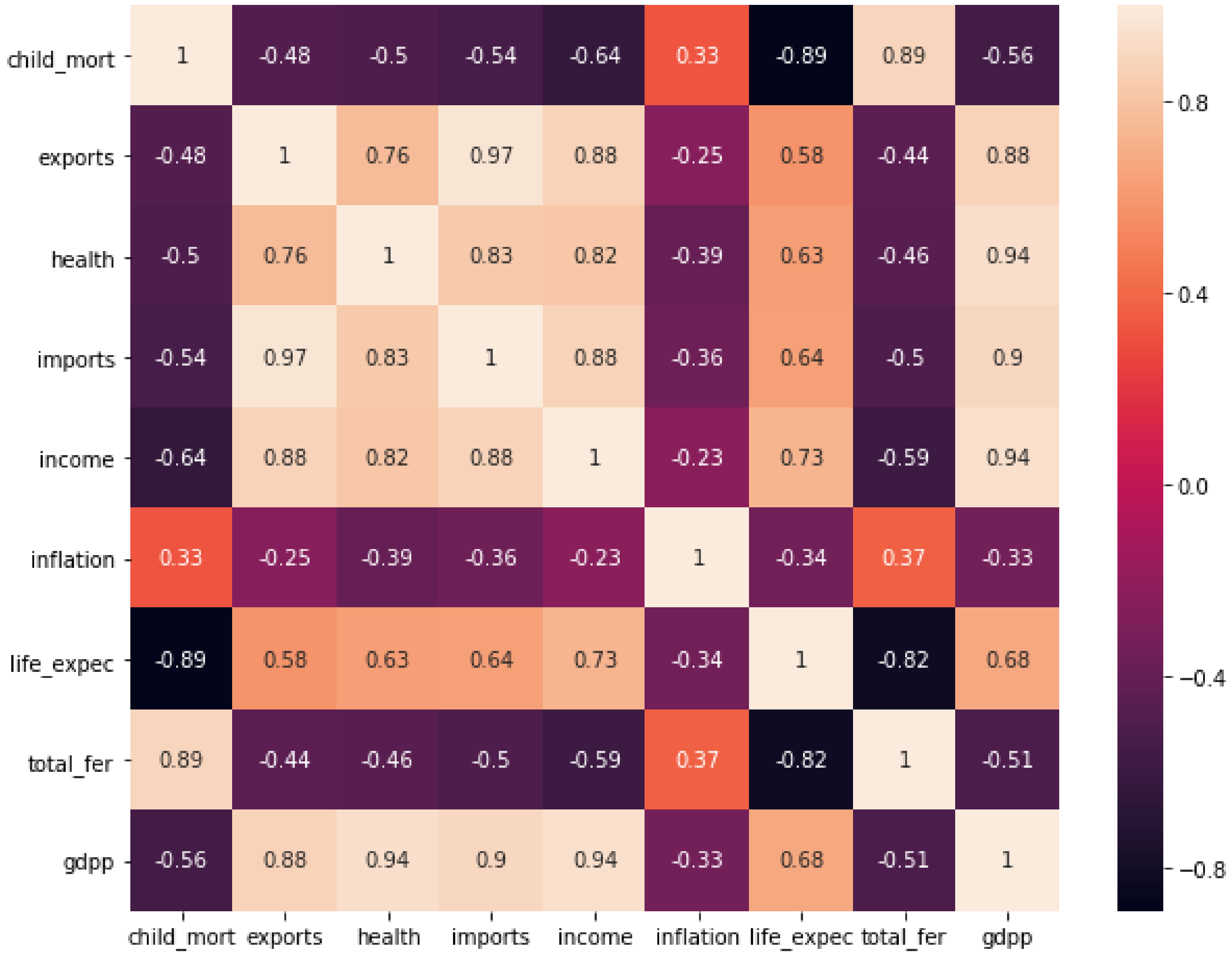
3.Countries with low income



4.Countries with low exports



- These simple visual checks speak a lot about the countries under analysis, but it does not help pin point suffering countries. Thus emphasizing the need for clustering.
- Majority of the underperforming countries are from Africa
- It is typically countries with very high infant mortality rates and low gdpp, that are the ones which require immediate help from the NGO.

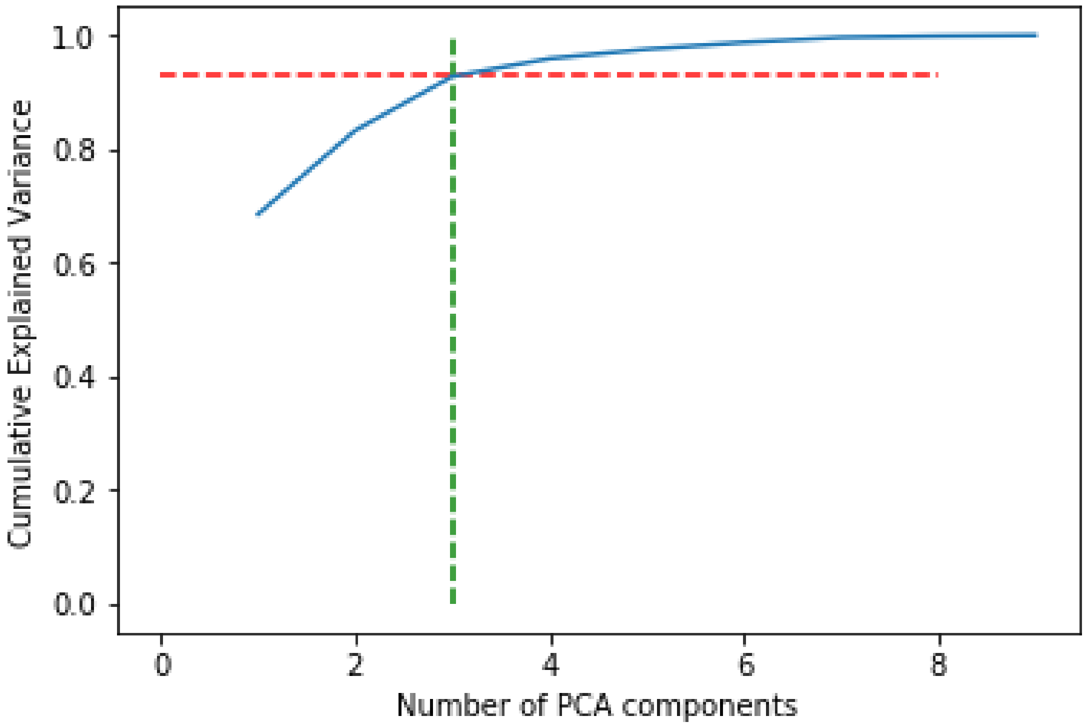


	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
child_mort	1	-0.48	-0.5	-0.54	-0.64	0.33	-0.89	0.89	-0.56
exports	-0.48	1	0.76	0.97	0.88	-0.25	0.58	-0.44	0.88
health	-0.5	0.76	1	0.83	0.82	-0.39	0.63	-0.46	0.94
imports	-0.54	0.97	0.83	1	0.88	-0.36	0.64	-0.5	0.9
income	-0.64	0.88	0.82	0.88	1	-0.23	0.73	-0.59	0.94
inflation	0.33	-0.25	-0.39	-0.36	-0.23	1	-0.34	0.37	-0.33
life_expec	-0.89	0.58	0.63	0.64	0.73	-0.34	1	-0.82	0.68
total_fer	0.89	-0.44	-0.46	-0.5	-0.59	0.37	-0.82	1	-0.51
gdpp	-0.56	0.88	0.94	0.9	0.94	-0.33	0.68	-0.51	1

From the **heatmap**, we can notice that all of the features have good correlation between them which can have a detrimental effect when building a clustering model.

This need for Dimensionality reduction is solved using **PCA** (Principal component Analysis).

Principal Component Analysis

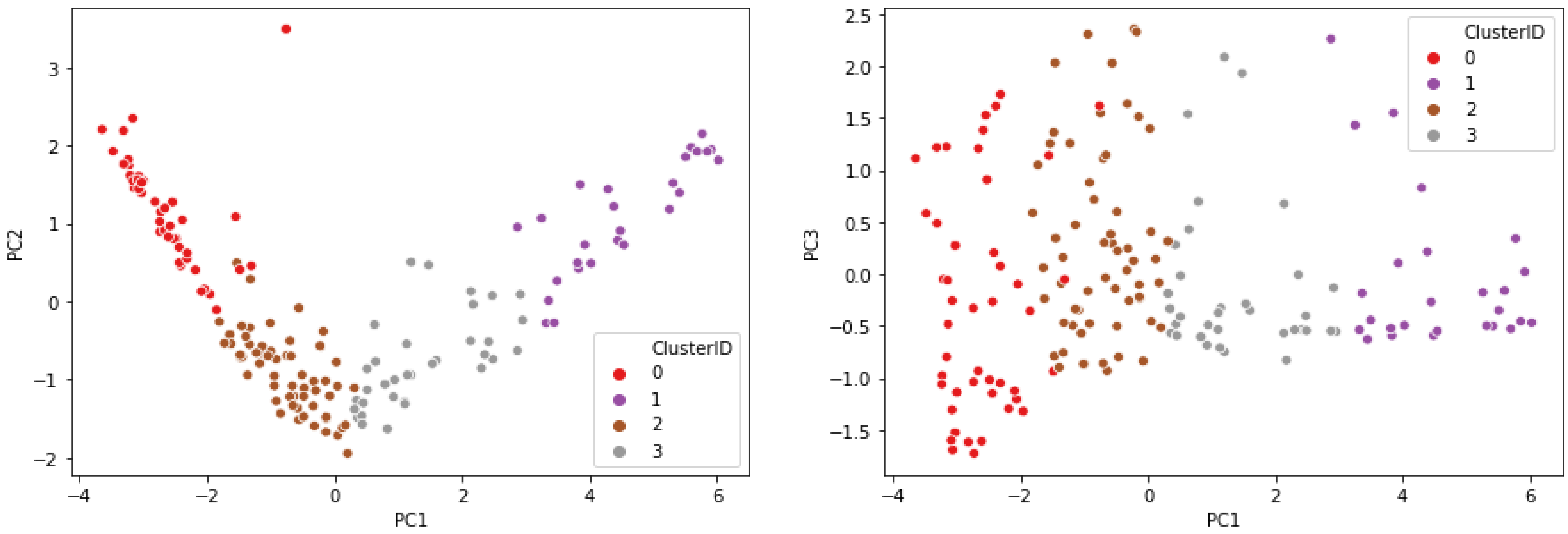


The **objective of PCA** is to find common factors, the so–called **principal components**, in form of linear combinations of the variables under investigation, and to rank them according to their importance.

From the **scree plot,** we can see that more than 90% of the variance can be captured by just 3 principal components, thus drastically cutting down the number of features required for our analysis.

Clustering

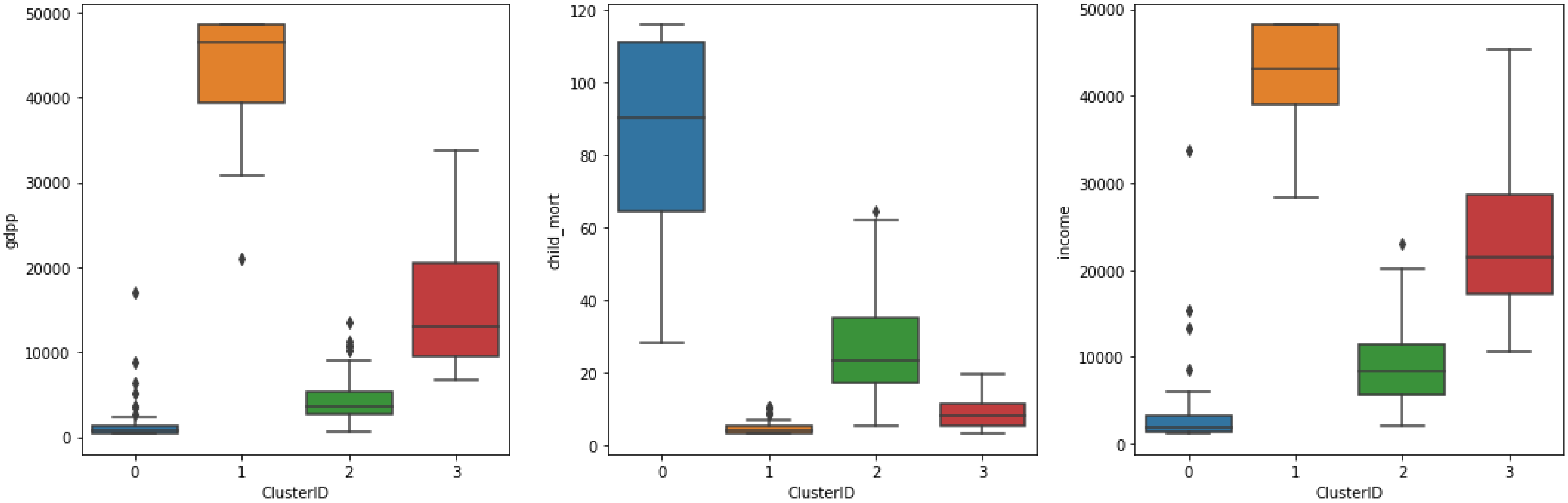
- **Clustering** is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups.
- Our aim is to cluster poor performing countries together and then use that cluster to identify countries that are in the direst need.
- This objective achieved using 2 clustering techniques:
  - **K-Means**
  - **Hierarchical clustering**
- Similar countries are clustered together by calculating the Euclidean distances between datapoints.
- K-Means is more popular, but in-order to run the algorithm, we have to decide on the number of clusters we want to create beforehand.
- Hierarchical clustering is more time consuming and machine intensive but we don’t need to know the number of clusters prior to execution.
- We make use of the best of both world’s to identify the top countries in need.



These scatterplots of principal components highlights how the clusters are formed using the K-means algorithm.

Few indicators of a good cluster are:

- Inter-cluster distances are less and Intra-cluster distances are more.
- Clear definitive differences between clusters.



- From the box-plots we could clearly deduce that countries in cluster 0 have the lowest income and gdpp but highest child mortality rates. Thus it is evident that these countries are the ones in direst need of aid.
- Cluster 3 countries are the best performing ones of the lot.

Conclusion

- From the initial Heatmaps and pair-plots, it was evident that the majority of features were heavily correlated and hence doesn't add additional information's which calls for the need of dimensionality reduction using PCA. By analyzing the clusters formed by K-means and Hierarchical clustering, I could pin point the countries which are performing the poorest; mainly by cross-checking how the clusters performed against predominant features such as **child mortality, income and gdpp.**

Countries the NGO should focus on helping are :

- Afghanistan
- Burkina Faso
- Burundi
- Central African Republic
- Congo, Dem. Rep.
- Guinea-Bissau
- Malawi
- Mozambique
- Niger
- Sierra Leone