PCA and CLUSTERING FOR HELP INTERNATIONAL

Problem Statement

- After the recent project that included a lot of awareness drives and funding programs, HELP International has been able to raise around $ 10 million.

- Now the need is 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.

Data Analysis

- We will begin the data analysis to find the which countries need aid by Principal Components Analysis. This will help us with the dimensionality reduction.
- Once PCA is complete, we group different countries into clusters using both K-Mean clustering and Hierarchical Clustering.
- We then identify the cluster that will have low socio economic and health factors.
- Finally, we will perform binning to get the final list of countries.

Principle Component Analysis

- We use the SVD randomized function and performed the analysis on PC1 and PC2.

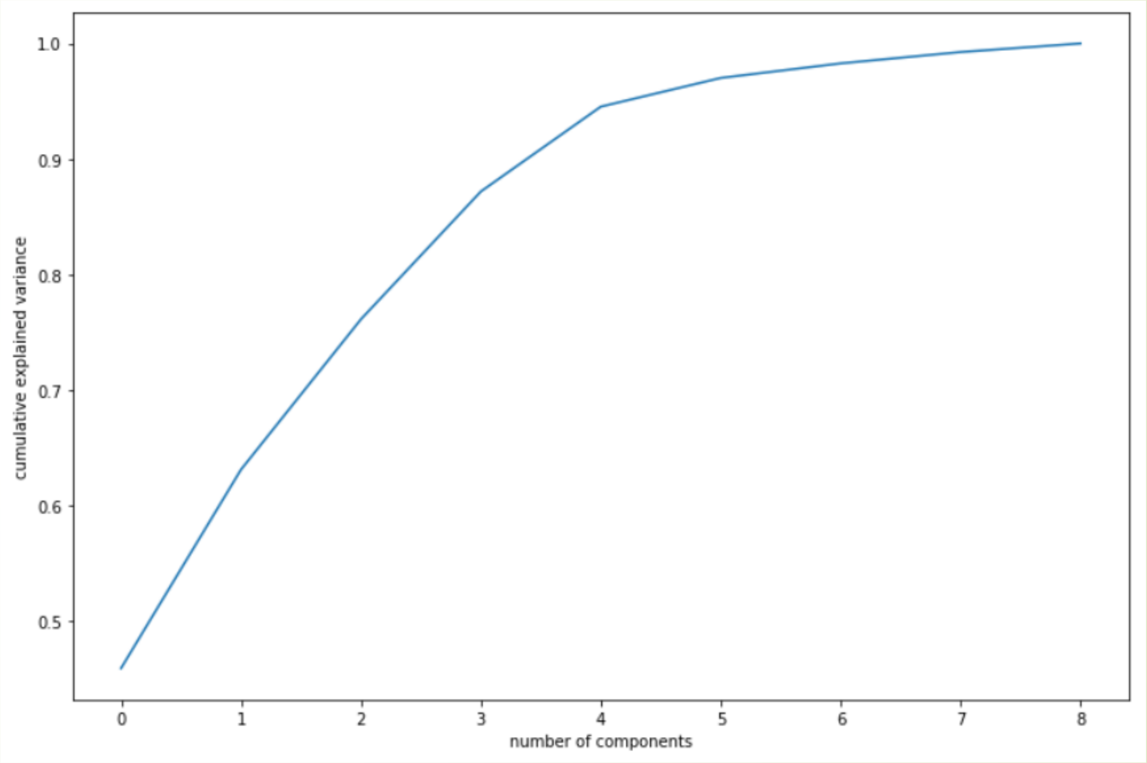
- We then use the Scree Plot to deduce that the number of components we require would be 4 i.e., we will have PC1, PC2, PC3 and PC4.

- We fit the data with these PCs and store it into a new data frame which will be used for clustering

PCA – scree plot

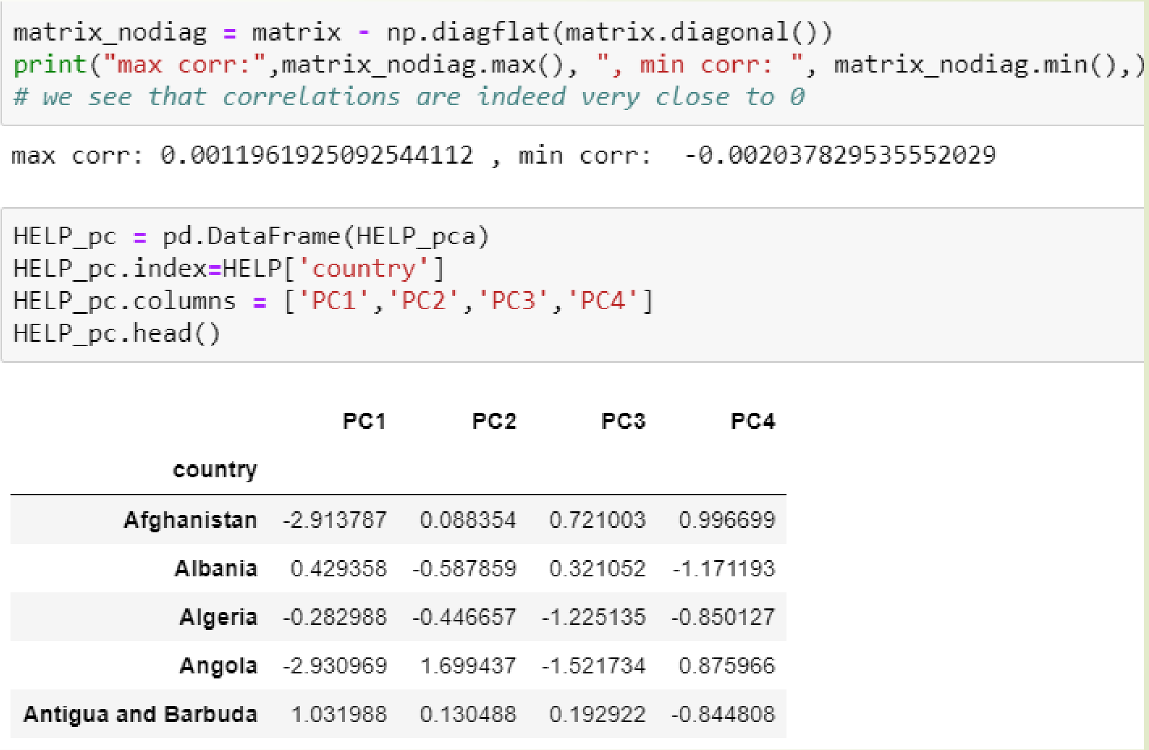
- The scree plot shows the variance of 95% at approximately 4 components.

- Hence the number of principal components we choose would be 4.



PCA – correlation matrix

- To ensure that the PCs have the least correlation, we check the minimum and maximum correlation between them.
- We see that the correlations is close to 0which lets us proceed.



```
matrix_nodiag = matrix - np.diagflat(matrix.diagonal())
print("max corr:",matrix_nodiag.max(), ", min corr: ", matrix_nodiag.min(),)
# we see that correlations are indeed very close to 0
```

```
max corr: 0.0011961925092544112 , min corr:  -0.002037829535552029
```

```
HELP_pc = pd.DataFrame(HELP_pca)
HELP_pc.index=HELP['country']
HELP_pc.columns = ['PC1','PC2','PC3','PC4']
HELP_pc.head()
```

	PC1	PC2	PC3	PC4
country				
Afghanistan	-2.913787	0.088354	0.721003	0.996699
Albania	0.429358	-0.587859	0.321052	-1.171193
Algeria	-0.282988	-0.446657	-1.225135	-0.850127
Angola	-2.930969	1.699437	-1.521734	0.875966
Antigua and Barbuda	1.031988	0.130488	0.192922	-0.844808

K Mean Clustering

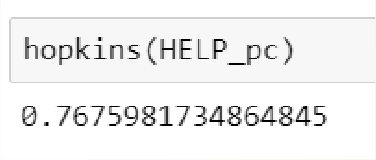
- We start with the HOPKINS Score to check the feasibility of forming the cluster using the Hopkins function. A score above 70% indicates it is feasible to form a good cluster. In our case, we get a score above 70% which is good to proceed.

- We then calculate Silhouette Score graph where the first peak would be considered for number of clusters. In our case, it shows 4.

- We run the sum of squared distances and plot the graph. Elbow of the curve indicates the number of clusters and in our case, elbow curve indicates to 4 as the number of clusters to be formed.

K Mean Clustering

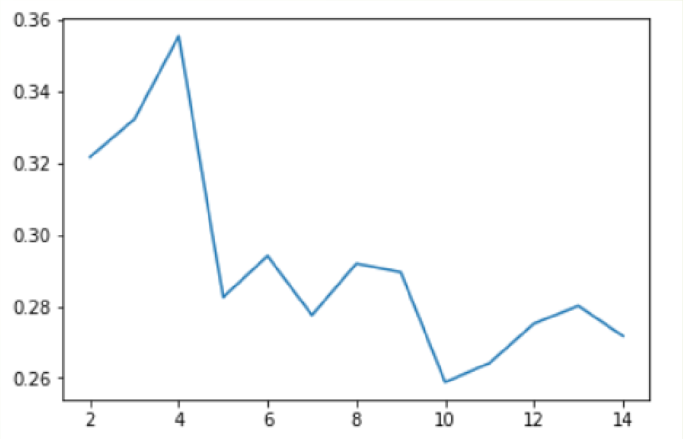
- Hopkins Score of 77%



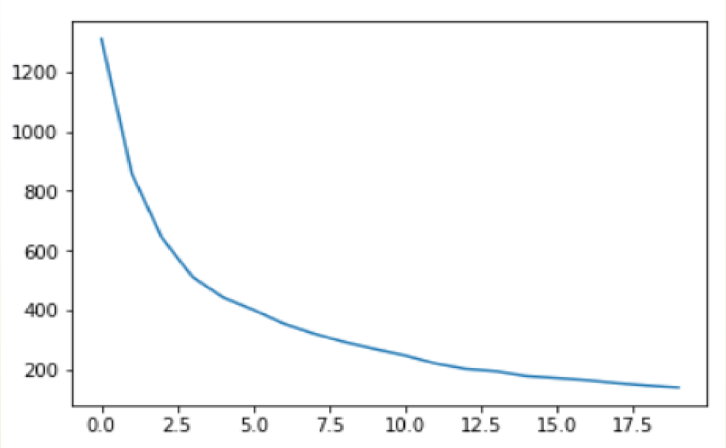
```
hopkins(HELP_pc)

0.7675981734864845
```

- Silhouette graph peaks to number 4



- Sum of Squares elbow curve corresponds to 4 clusters



K Mean Clustering Visualization

- We create a new data frame where each column will have their mean cluster-wise.

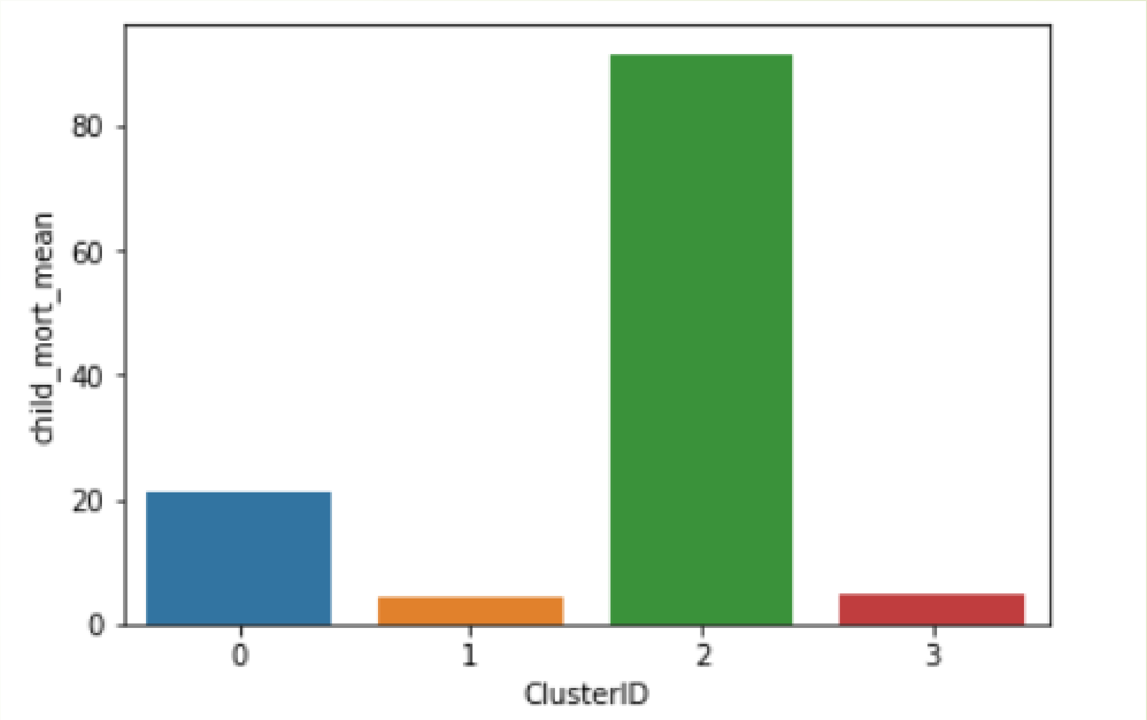
- We then plot graphs where the clusters are plotted against each of the column for data analysis.

- We then deduce that Child Mortality, Income, Inflation and GDPP as the key factors for us to identify the poor countries that need direst aid.

- We can then see that Cluster 2 is the one with high child mortality, low income, high inflation and low GDPP.

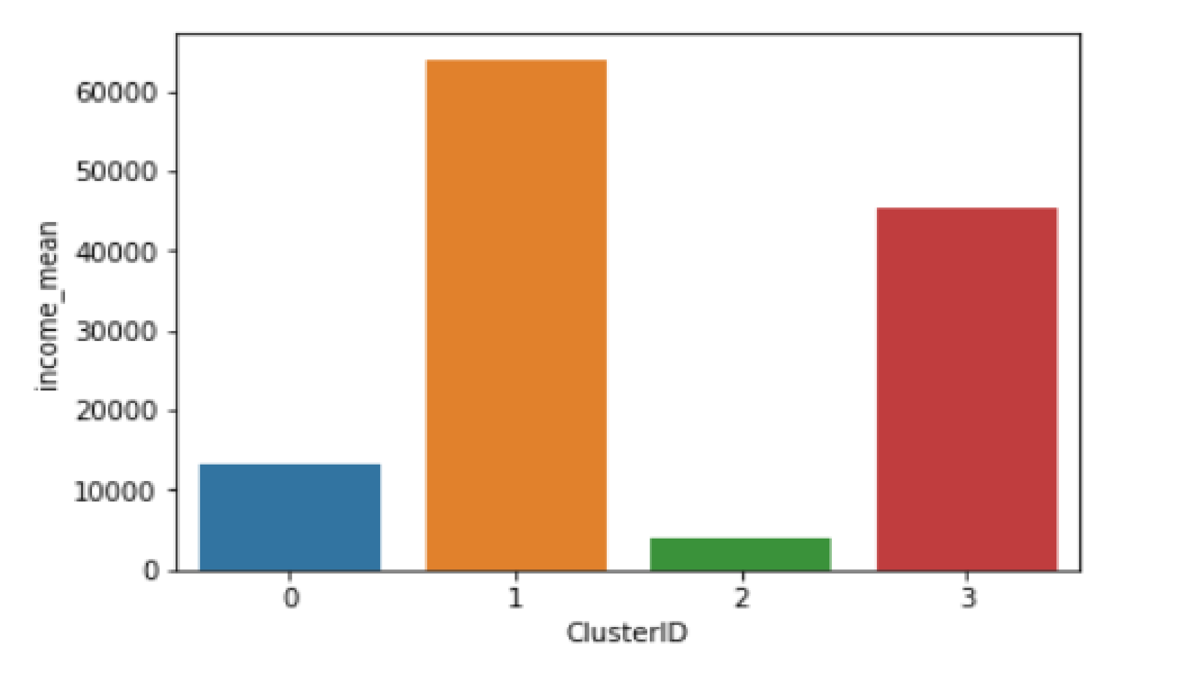
Child Mortality

- Cluster 2 shows high child mortality, above 80%



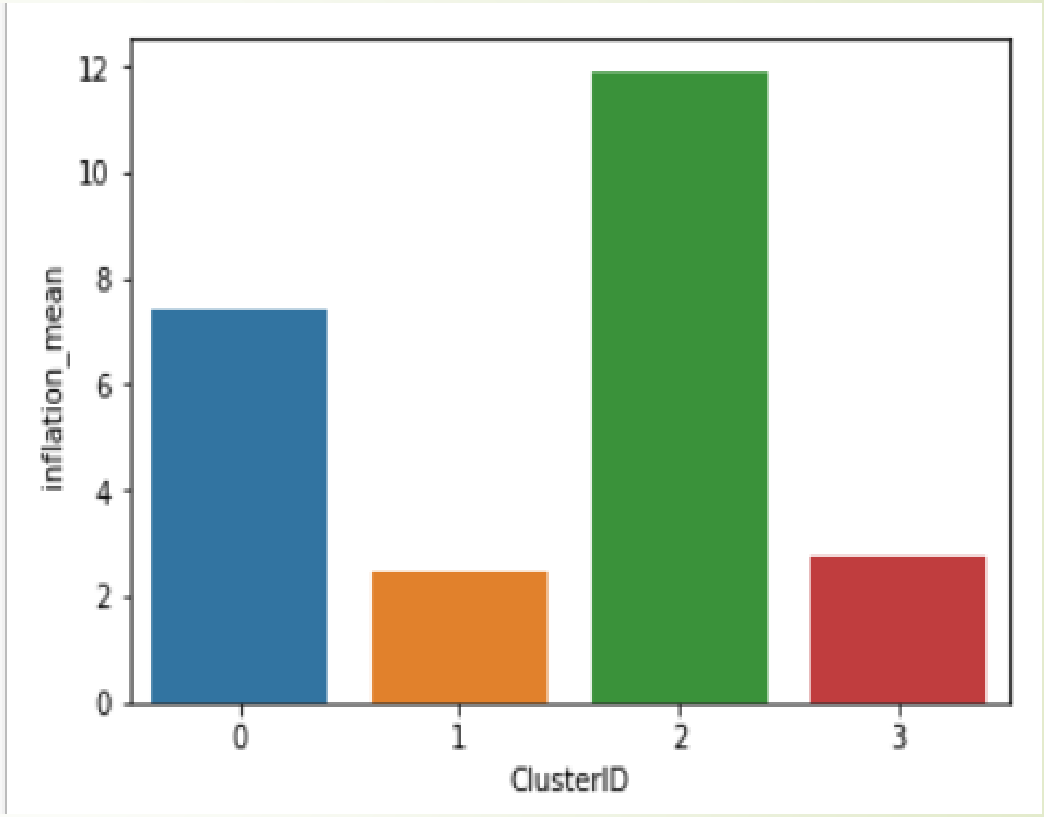
Income

- Cluster 2 shows low income rate, less than 4000.



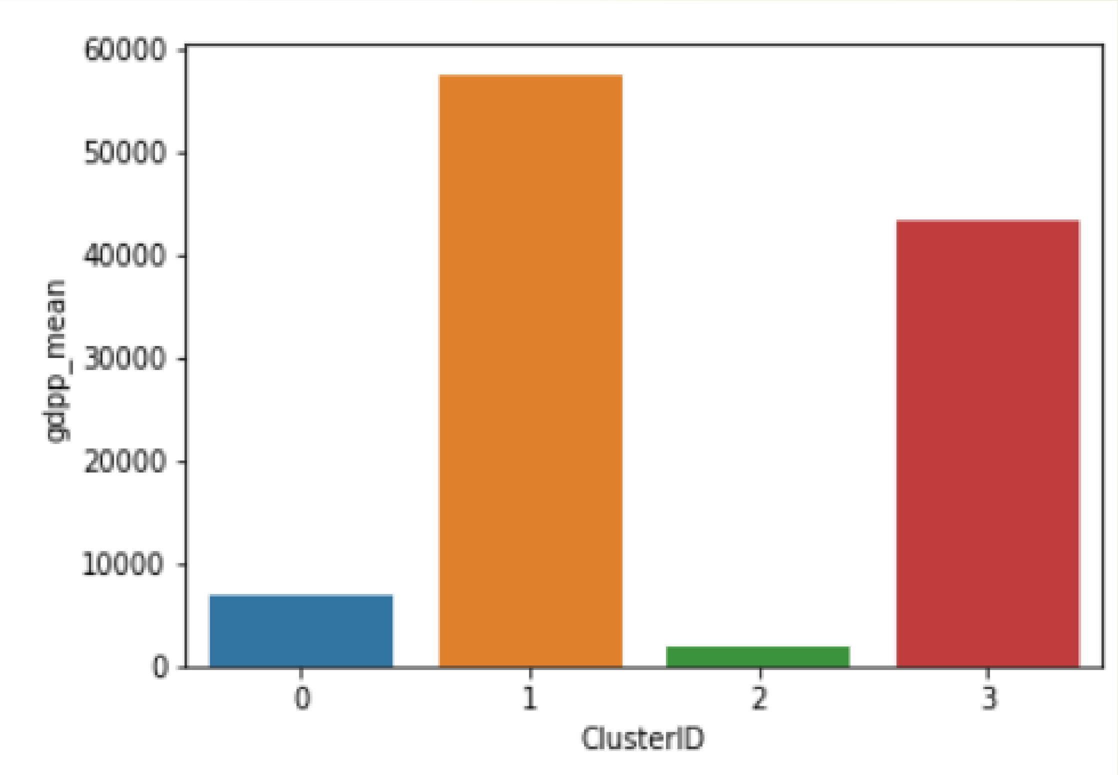
Inflation

- Cluster 2 shows high inflation, around 12%.



GDPP

- Cluster 2 shows low GDPP.

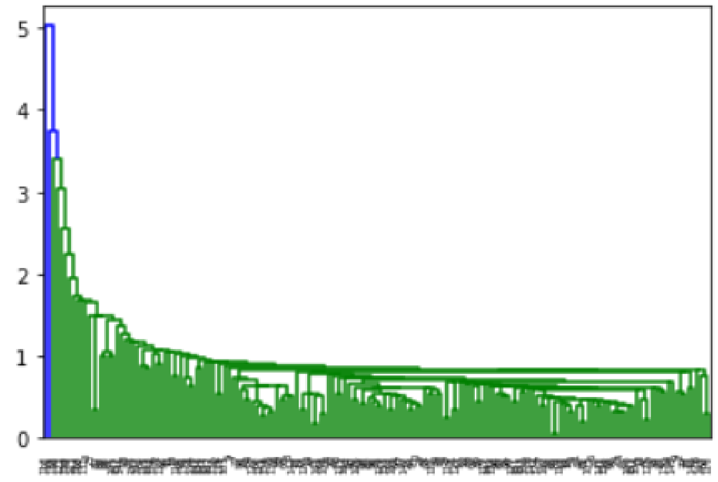


Hierarchical Clustering

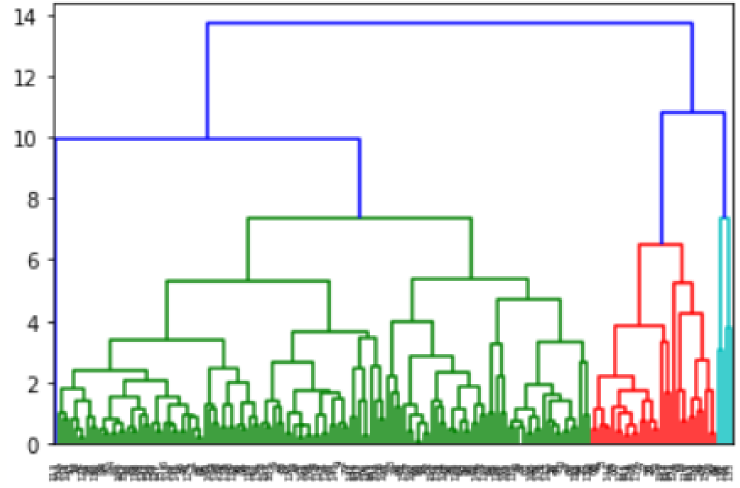
- We create a dendrogram using single linkage and complete linkage.
- Since its huge data set, we will decide on a height to cut the dendrogram
- After few trail and error, we decide on the height to be 3.
- We create a new data frame where each column will have their mean cluster-wise.
- We then plot graphs where the clusters are plotted against each of the column for data analysis.
- We then deduce that Child Mortality, Income, Inflation and GDPP as the key factors for us to identify the poor countries that need direct aid.
- We can then see that Cluster 0 is the one with high child mortality, low income, high inflation and low GDPP.

Hierarchical Clustering Visualization

- Single Linkage

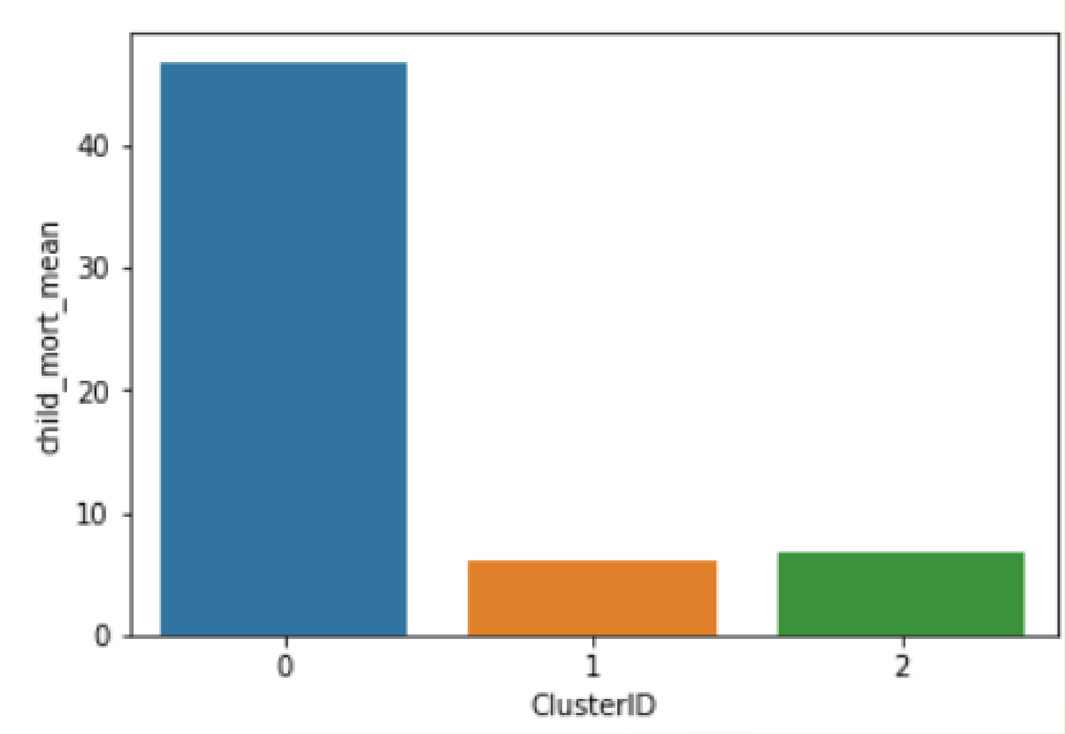


- Complete Linkage



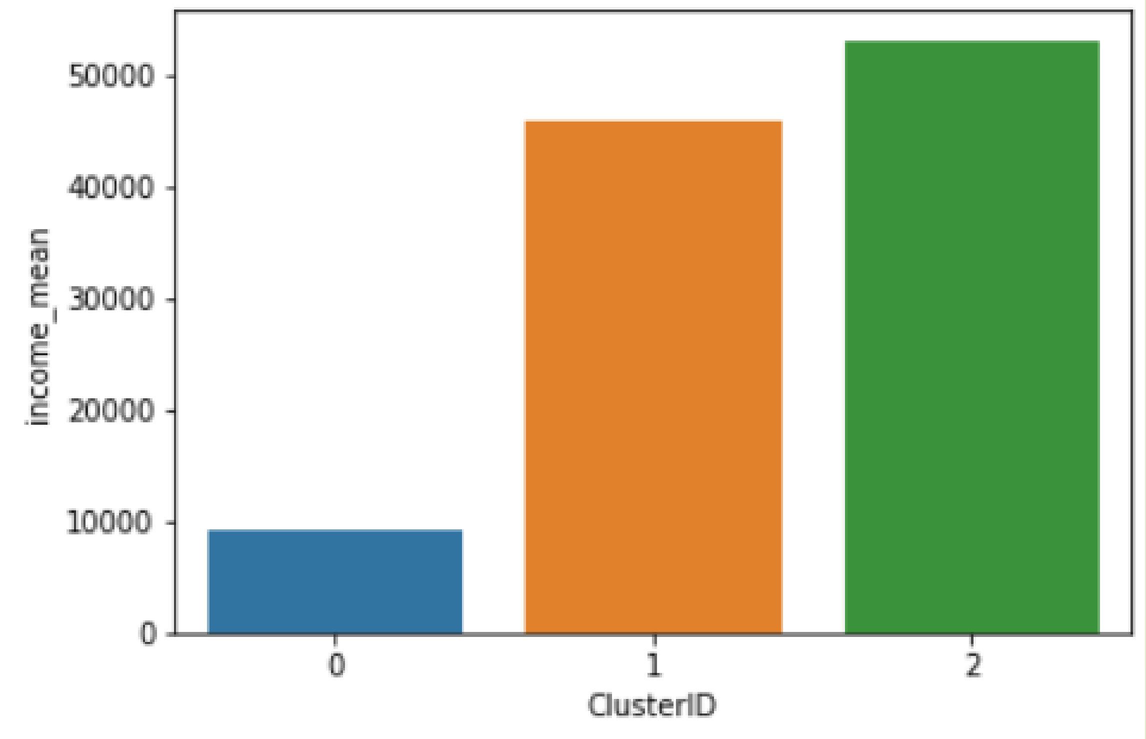
Child Mortality

- Cluster 0 shows high child mortality, above 40%



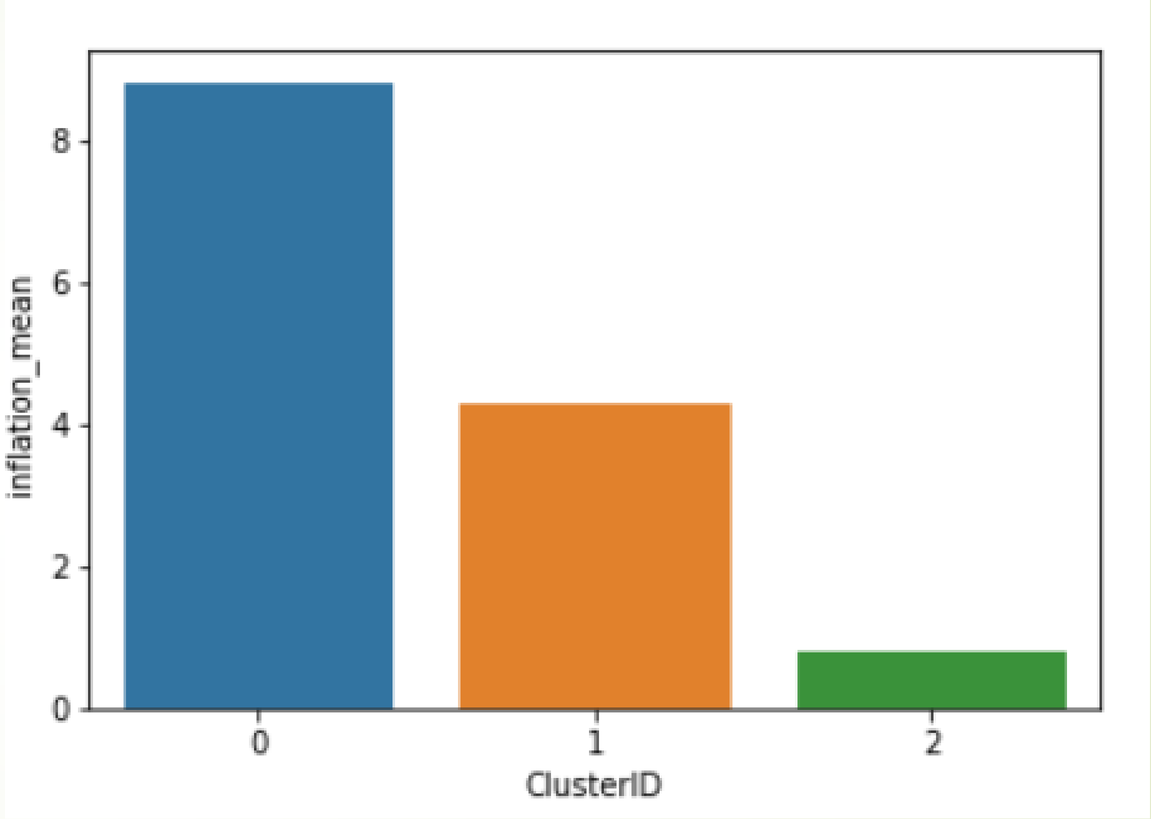
Income

- Cluster 0 shows low income rate, less than 7000.



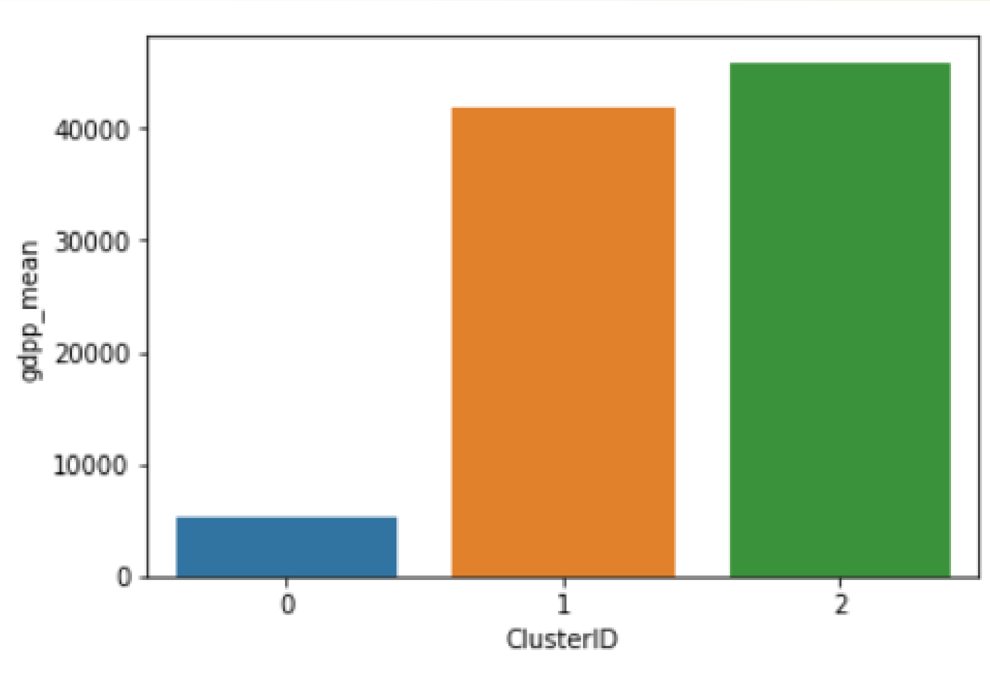
Inflation

- Cluster 0 shows high inflation, above 8%.



GDPP

- Cluster 0 shows low GDPP.



Summary

- We see that Cluster 2 from K Mean clustering and Cluster 0 from Hierarchical Clustering comprises of those countries that would need direst aid.
- The K Mean clustering gives a list of 48 countries and the Hierarchical Cluster gives a list of 132 countries.
- We go with the list of countries from K Mean to get the final list of countries. As 10 million would be less to be distributed among 48 countries, we do the below steps to narrow down the list of countries.
- We can take **the mean** of Child Mortality, Income, Inflation and GDPP forthat particular cluster i.e., Cluster 2 and set it to a new data frame

Summary

- We consider the above mean value of each key factor as a cut-off to get the set of countries that fall within that cut-off value.
- For child mortality we choose > 91%, income we choose < 3898, inflation we choose >11.9%, gdpp we choose < 1910.
- We then get a list of five countries that would require aid.
- The countries that would require aid would be,
  - Burundi
  - Democratic Republic of Congo
  - Guinea
  - Mauritania
  - Sierra Leone