

# CLUSTERING AND PCA ASSIGNMENT

Done By : Meenu Jomi

# 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. It runs a lot of operational projects from time to time along with advocacy drives to raise awareness as well as for funding purposes.

After the recent funding programmes, 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. The significant issues that come while making this decision are mostly related to choosing the countries that are in the direst need of aid.

# OBJECTIVE

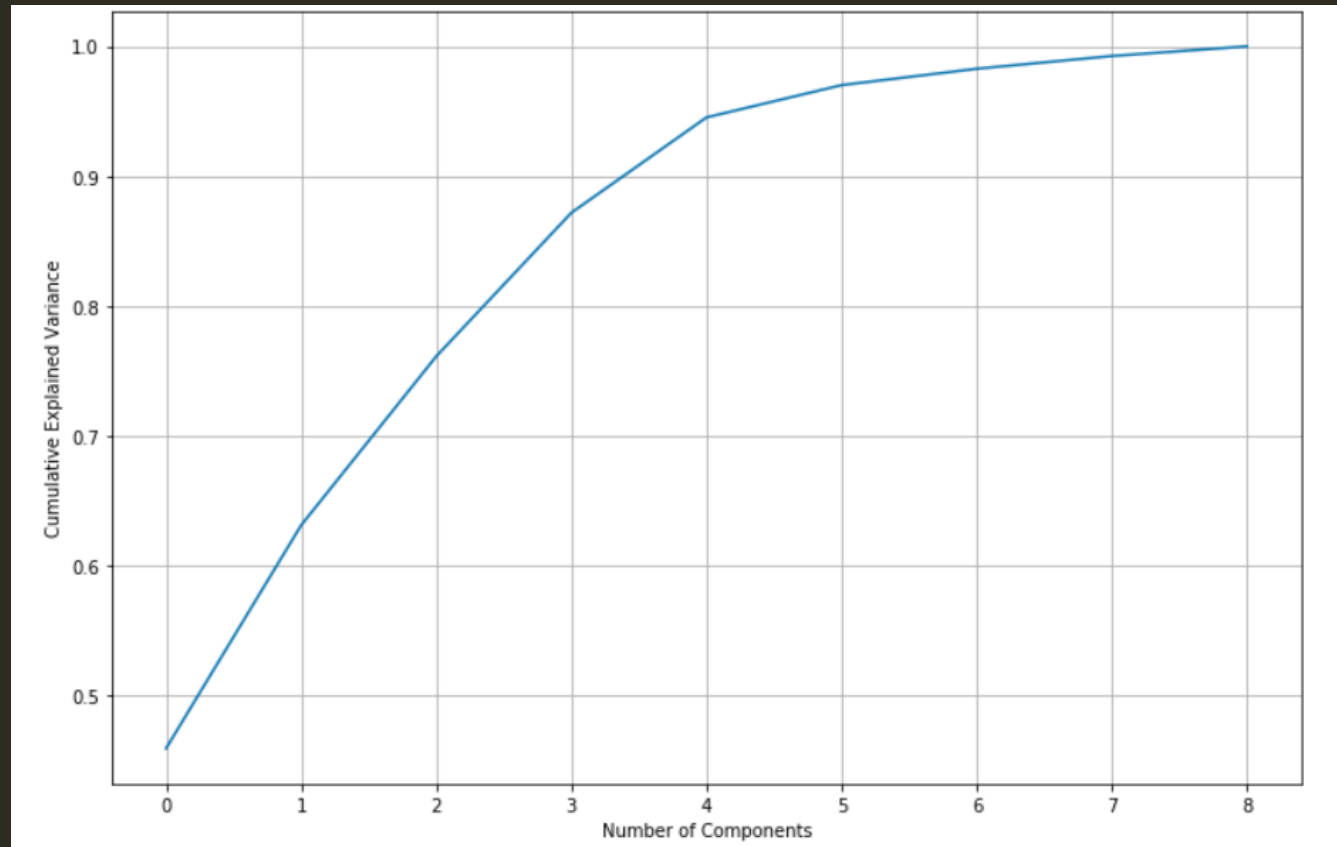
The requisite is:

- To categorise the countries using some socio-economic and health factors that determine the overall development of the country.
- To suggest the countries which the CEO needs to focus on the most.

# DATA PROCESSING

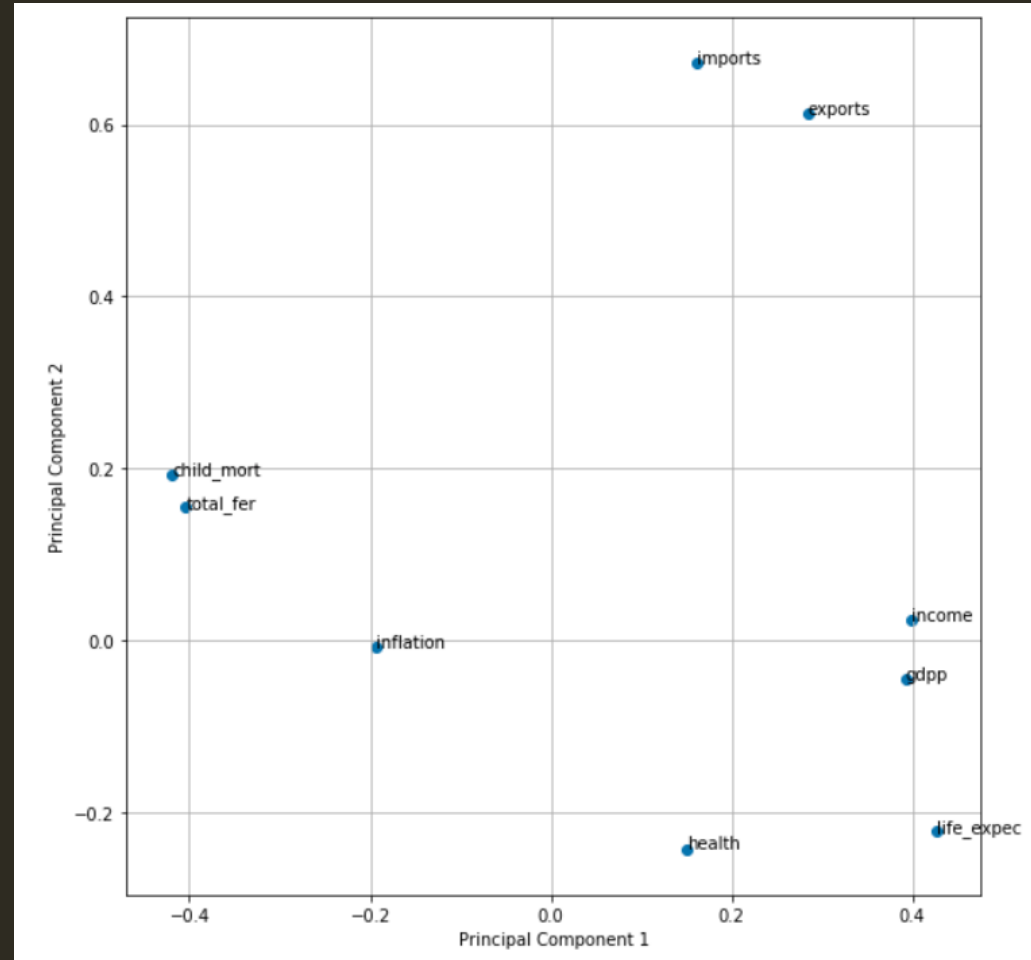
- It was found that there were no null values
- There were also no duplicate values for country
- There were a few outliers and they were treated later on during PCA
- The data was standardized for Principal Component Analysis

# PCA (SCREEPLOT)



**4 components are good enough to get a 95% of variance in the data. So PC is selected to be 4**

# PCA



# CLUSTERING

- Both the methods K means and Hierarchical Clustering was used on the 4 PCA components
- For K means ,  $K = 3$  was taken.
- While doing the Hopkins Statistics a value of 0.77 was attained.

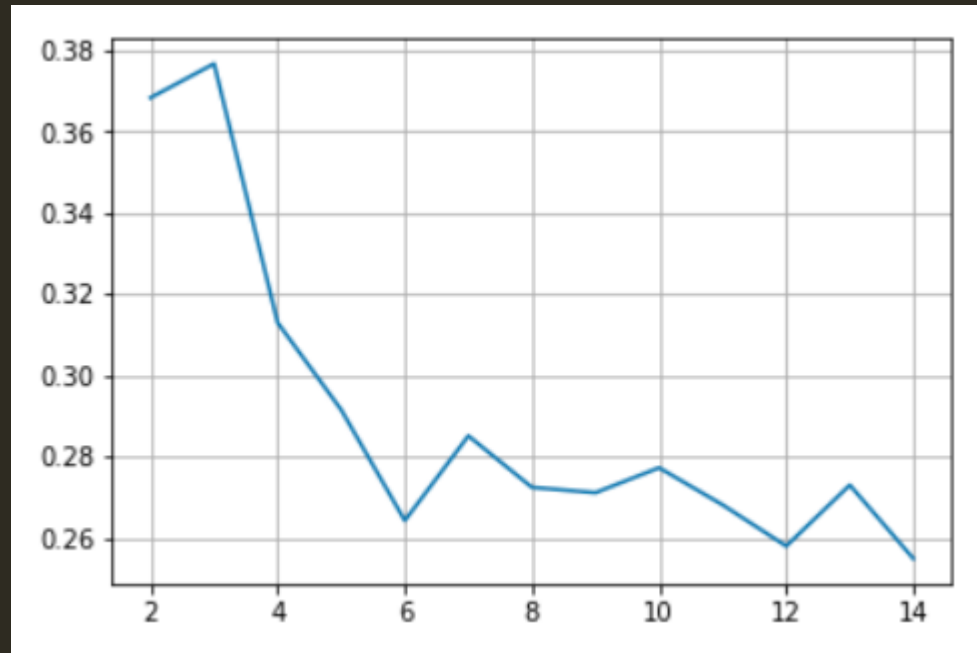
[ If the values are:

0.01 - 0.3 : Low chance of clustering

around 0.5 : Random

0.7 - 0.99 : High chance of clustering]

# SILHOUETTE ANALYSIS (PEAKING AT 3)



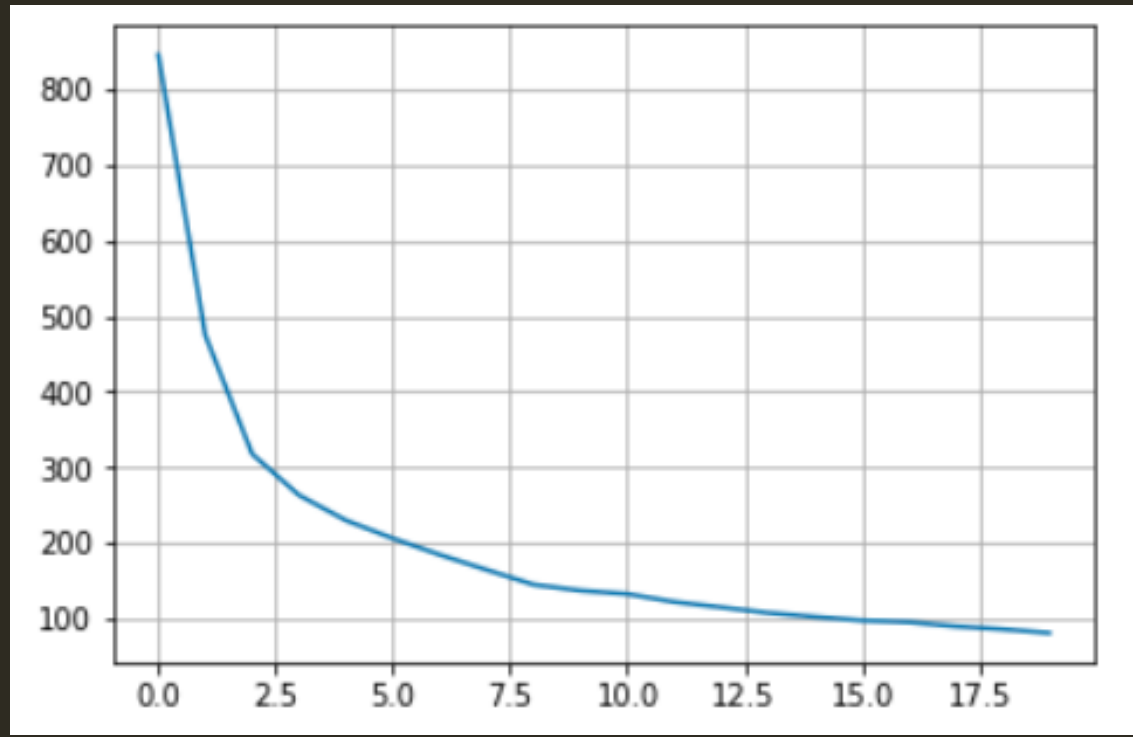
**Value of the silhouette score range is in between -1 to 1.**

**A score closer to 1 : The data point is very similar to other data points in the cluster**

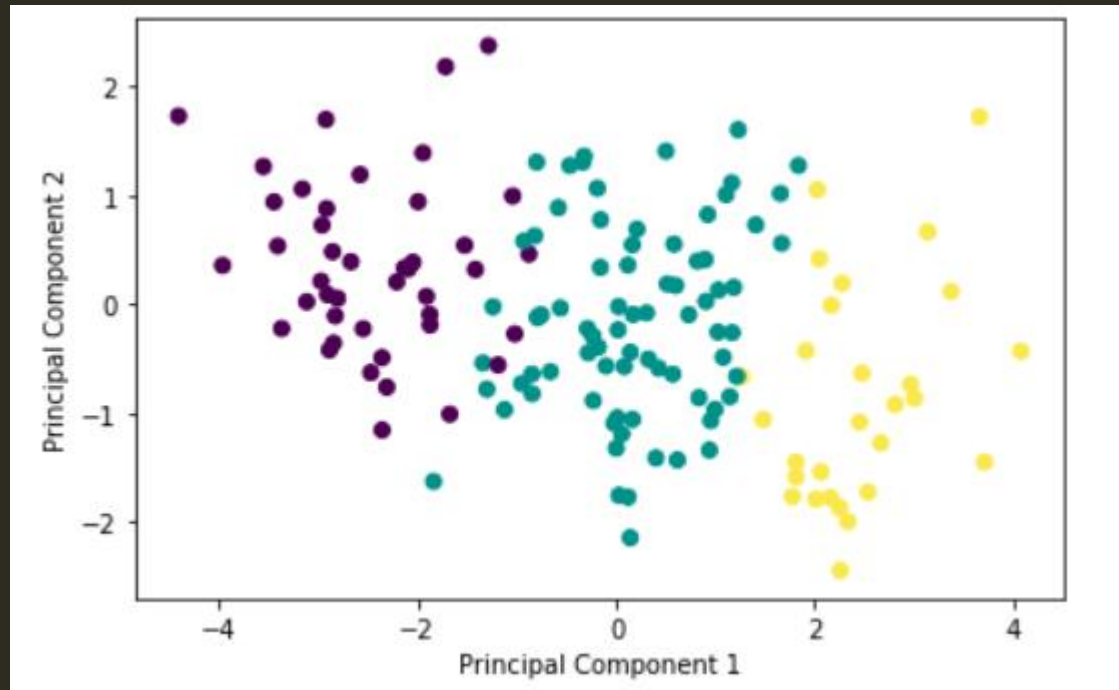
**A score closer to -1 : The data point is not similar to the data points in its cluster.**



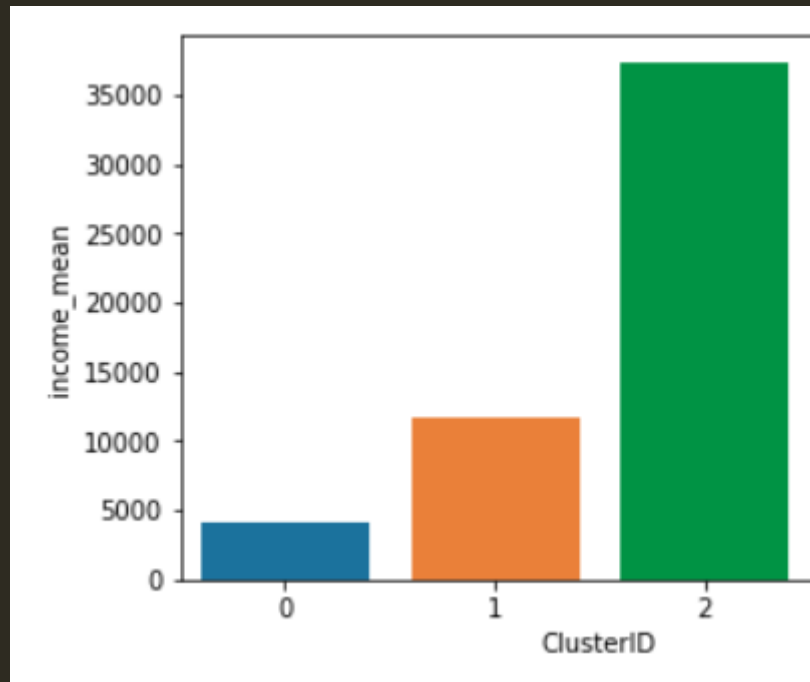
# SUM OF SQUARED DISTANCES



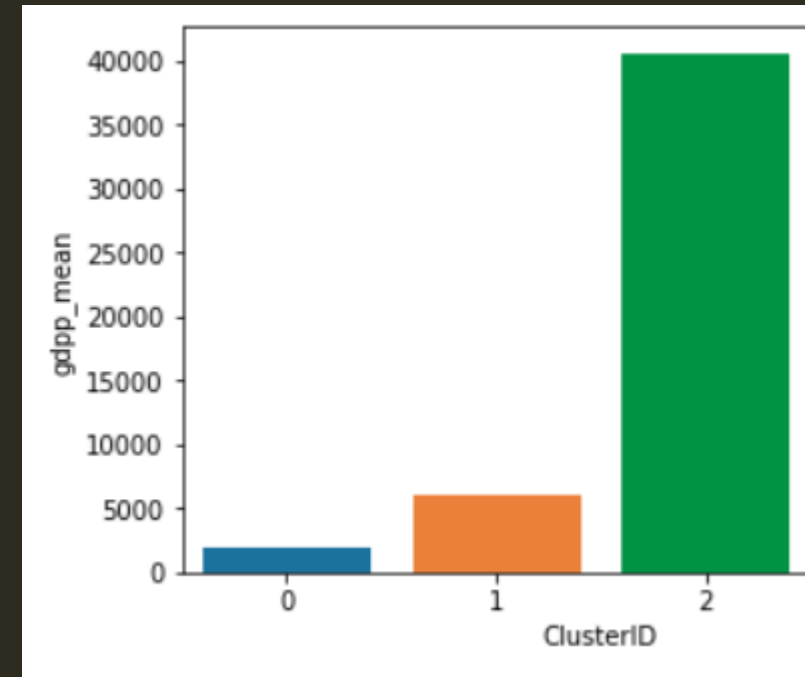
# CLUSTERING



# CLUSTERING

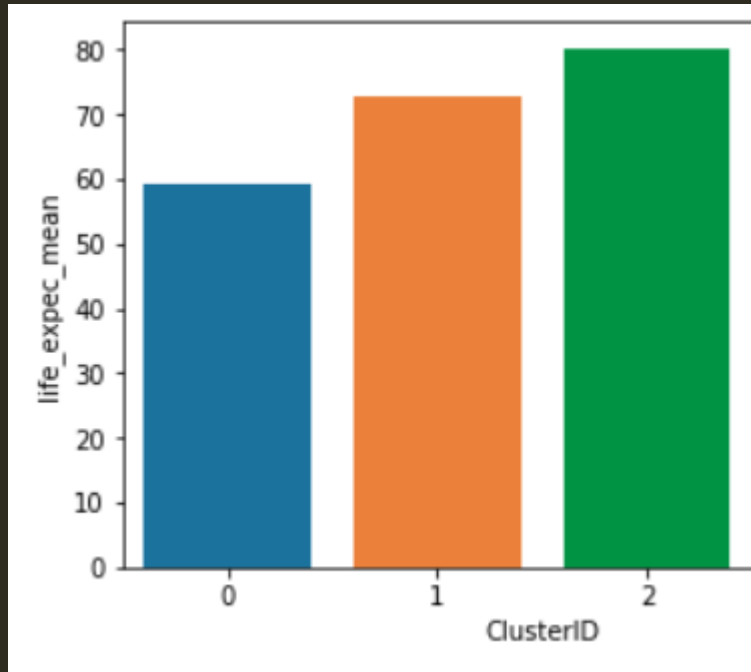


**Net income per person**

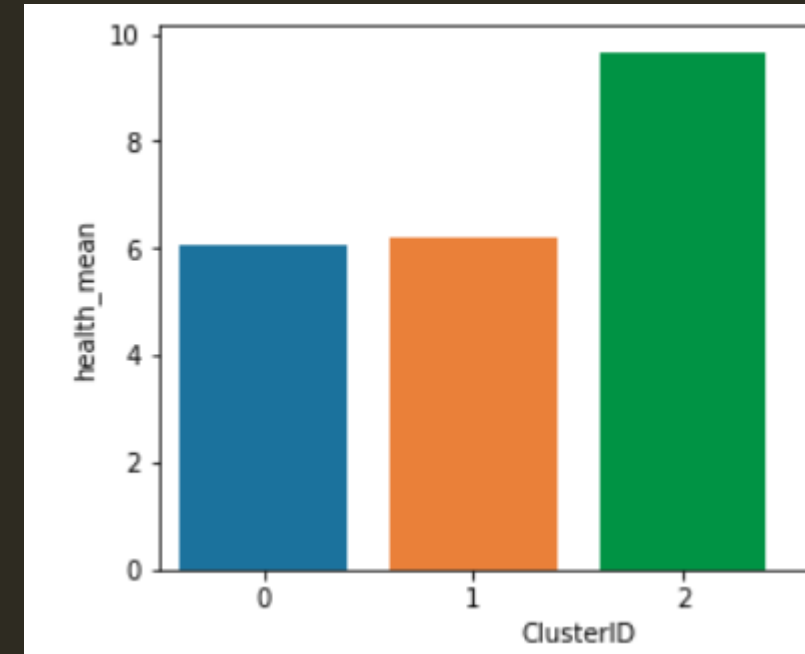


**The GDP per capita**

# CLUSTERING

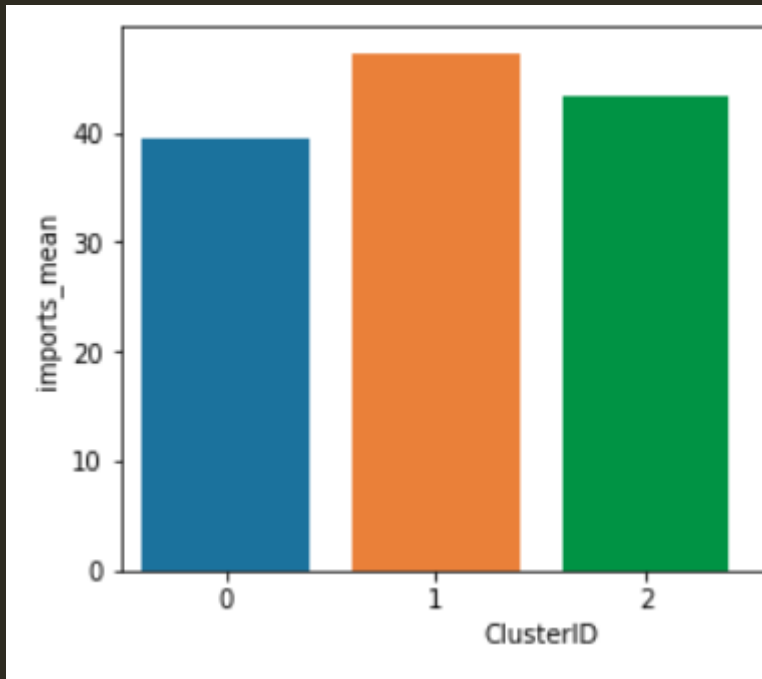


**The average number of  
years a new born child  
would live**

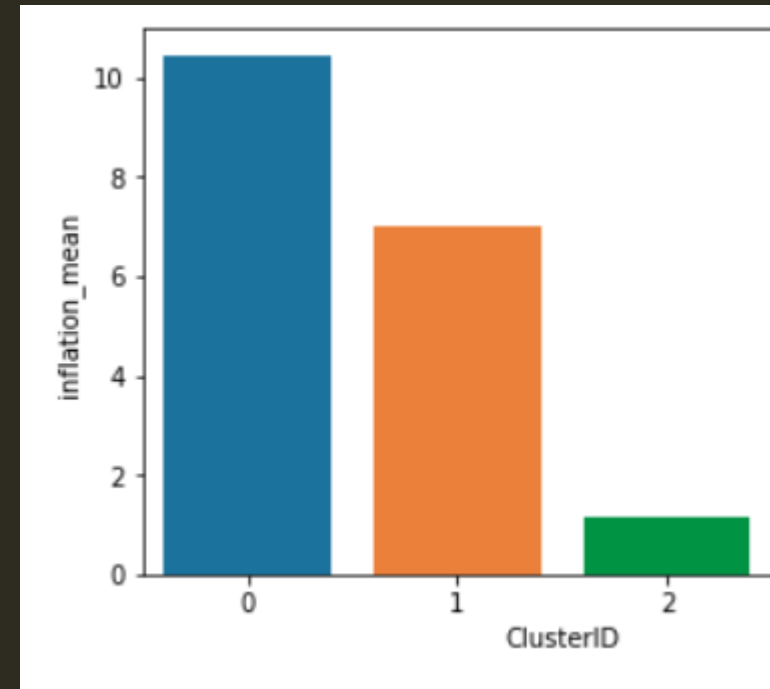


**Total health spending**

# CLUSTERING

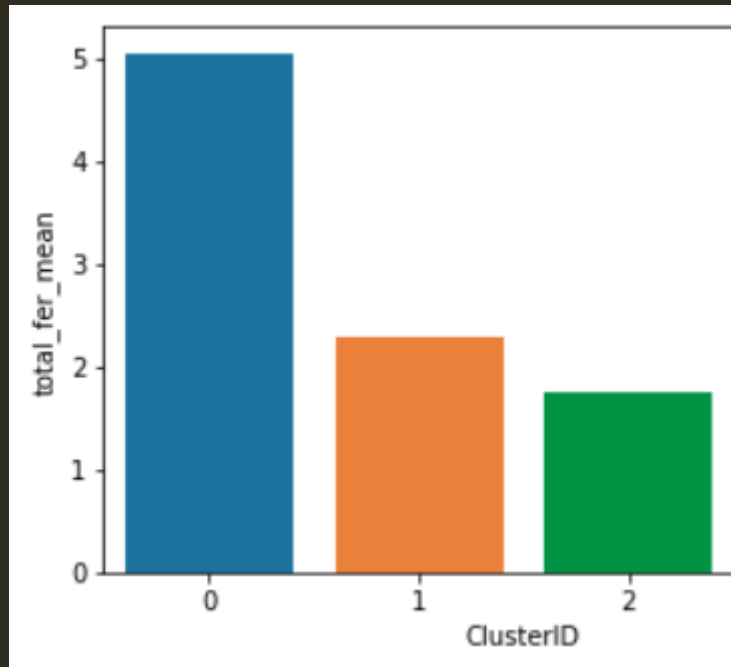


**Imports of goods and services**

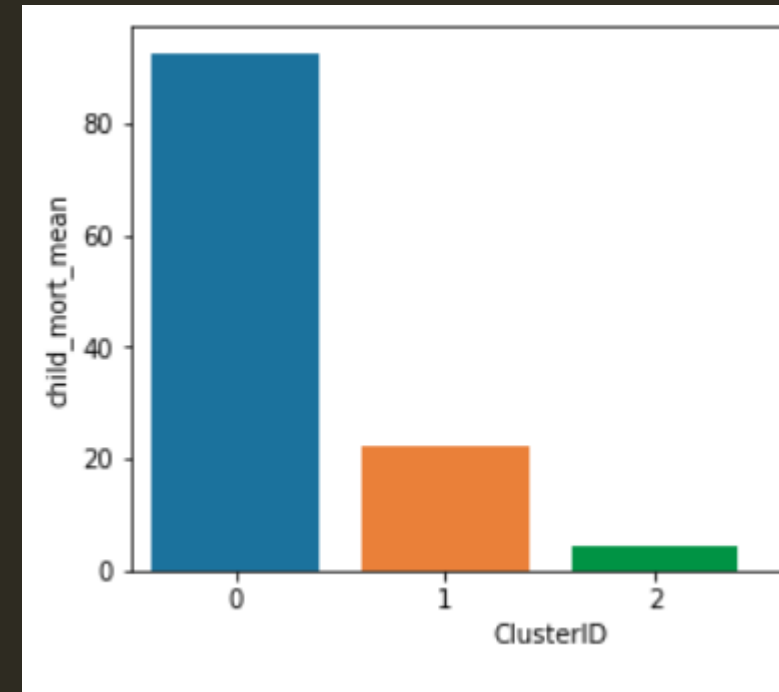


**The measurement of the annual growth rate**

# CLUSTERING

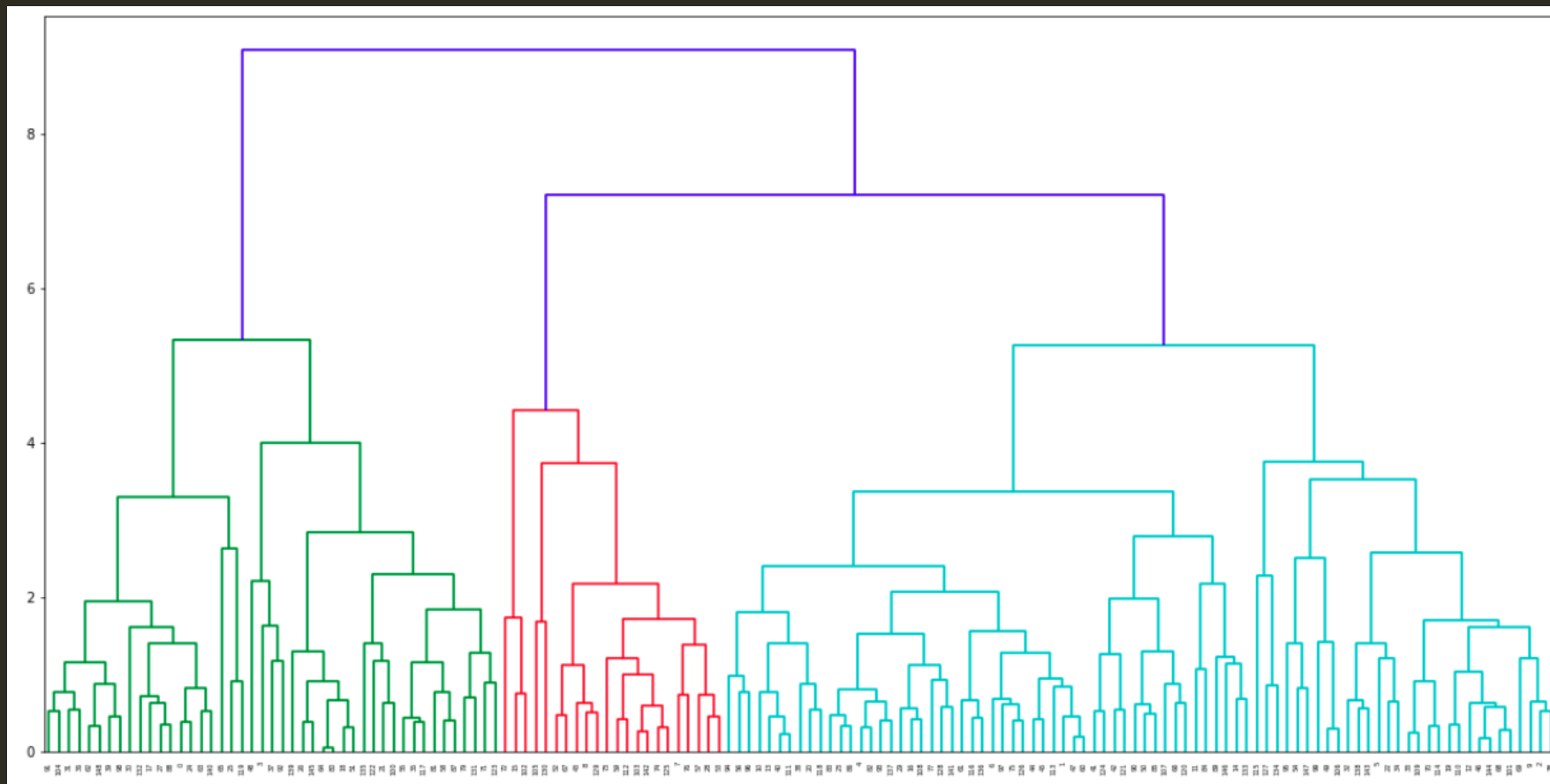


**The number of children  
that would be born**



**Child mortality rate**

# HIERARCHICAL CLUSTERING



**This shows how many clusters the data can be split into**

# CONCLUSION

The countries that require help the most are listed below:

Afghanistan, Angola, Benin, Botswana, Burkina Faso, Burundi, Cameroon, Central African Republic, Chad, Comoros, Congo, Dem. Rep., Congo, Rep., Cote d'Ivoire, Equatorial Guinea, Eritrea, Gabon, Gambia, Ghana, Guinea, Guinea-Bissau, Haiti, Iraq, Kenya, Lao, Madagascar, Malawi, Mali, Mauritania, Mozambique, Namibia, Niger, Pakistan, Rwanda, Senegal, Sierra Leone, South Africa, Sudan, Tanzania, Timor-Leste, Togo, Uganda, Yemen and Zambia.

These countries have

- very low rate of net income per person, GDP per capita, average number of years a new born child would live, total health spending and imports of goods and services.
- very high rate of measurement of the annual growth rate, number of children that would be born and child mortality rate.

It is clear that these countries require very quick aid in terms of money, education and services.