# Assignment: Clustering and PCA

SMRUTI RANJAN PARIDA

# Problem Statement

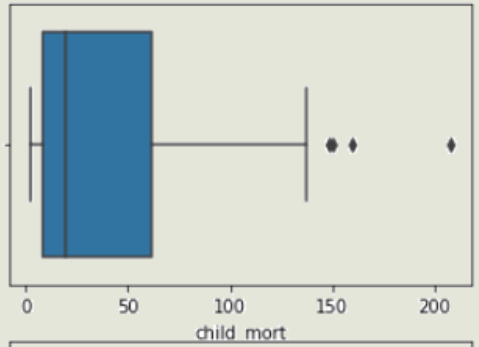
- An International humanitarian NGO that is committed to fighting poverty and providing the people of backward countries with basic amenities and relief during the time of disasters and natural calamities. It runs a lot of operational projects from time to time along with advocacy drives to raise awareness as well as for funding purposes.
- After the recent funding programmes, they have been able to raise around 10 million USD. 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.
- The Objective is to cluster the countries by the socio-economic and health factors and determine the overall development of the country.
- Available data :
  - datasets containing those socio-economic factors.

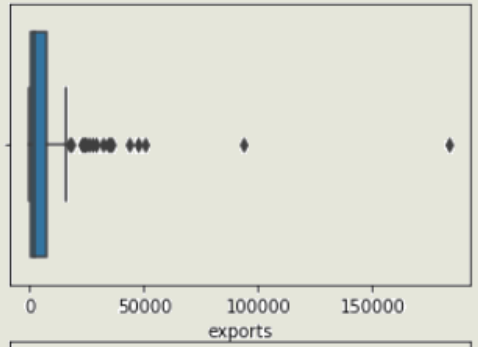
# Summary of Data set

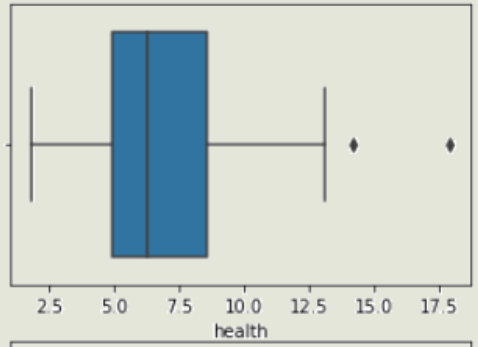
| | child_mort | exports | health | imports | income | inflation | life_expec | total_fer | gdpp |
|---|---|---|---|---|---|---|---|---|---|
| **count** | 167 | 167 | 167 | 167 | 167 | 167 | 167 | 167 | 167 |
| **mean** | 38.27 | 41.11 | 6.82 | 46.89 | 17144.69 | 7.78 | 70.56 | 2.95 | 12964.16 |
| **std** | 40.33 | 27.41 | 2.75 | 24.21 | 19278.07 | 10.57 | 8.89 | 1.51 | 18328.7 |
| **min** | 2.6 | 0.11 | 1.81 | 0.07 | 609 | -4.21 | 32.1 | 1.15 | 231 |
| **25%** | 8.25 | 23.8 | 4.92 | 30.2 | 3355 | 1.81 | 65.3 | 1.8 | 1330 |
| **50%** | 19.3 | 35 | 6.32 | 43.3 | 9960 | 5.39 | 73.1 | 2.41 | 4660 |
| **75%** | 62.1 | 51.35 | 8.6 | 58.75 | 22800 | 10.75 | 76.8 | 3.88 | 14050 |
| **max** | 208 | 200 | 17.9 | 174 | 125000 | 104 | 82.8 | 7.49 | 105000 |

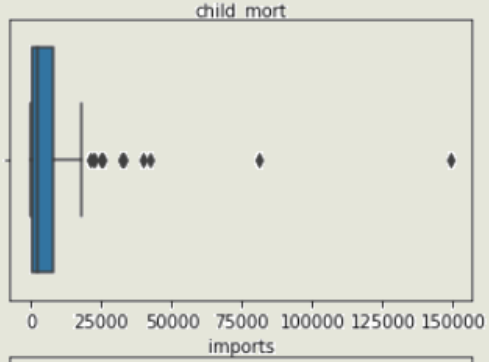
No of Countries for which data is available : 167
No of columns in the dataset : 10
No of Missing values : 0
Outlier Treatment is required for the columns like Income, gdpp, inflation

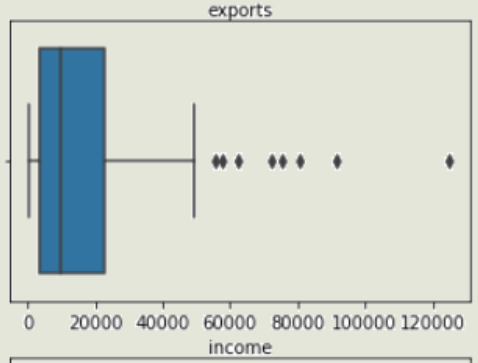
# Outliers in Data

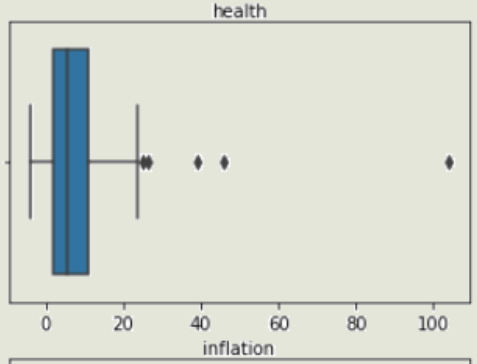


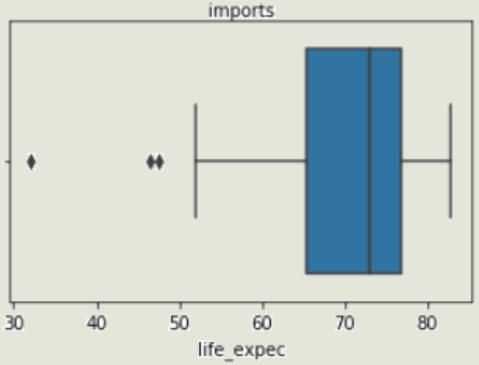



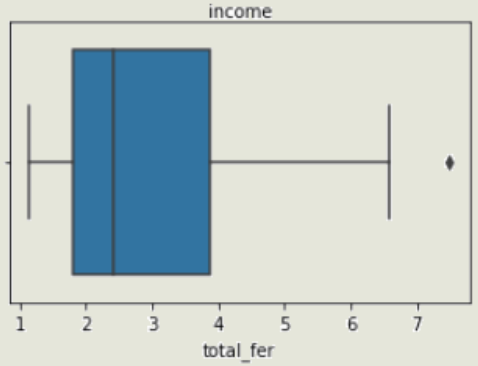



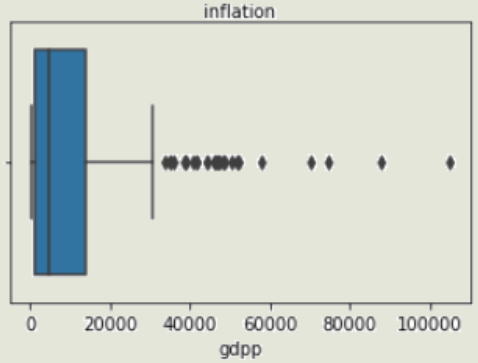



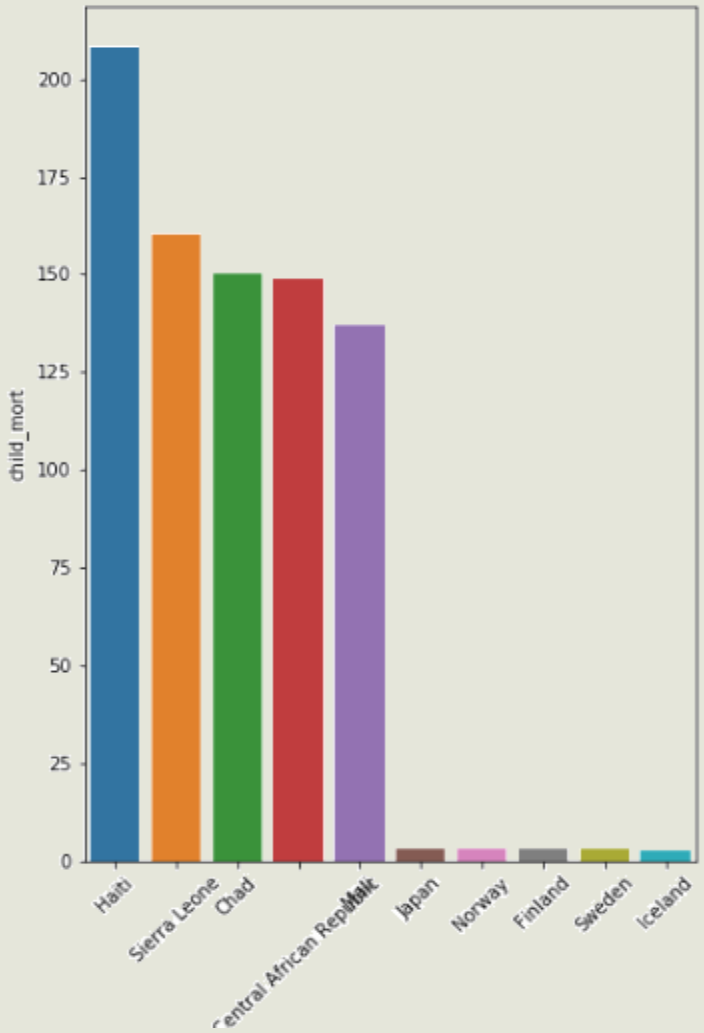



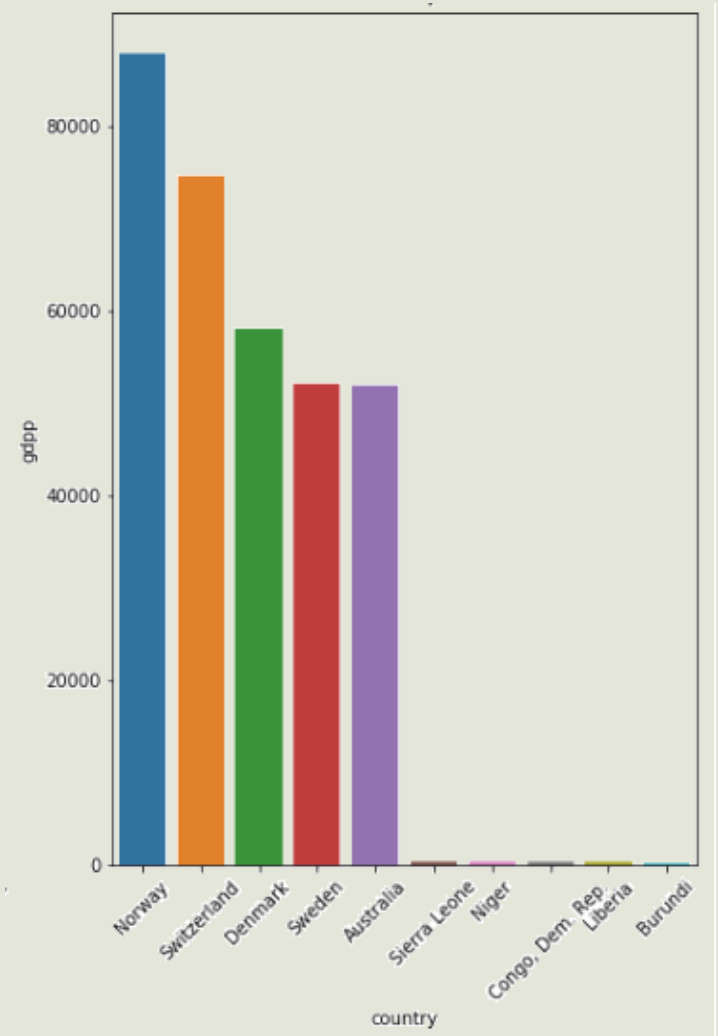



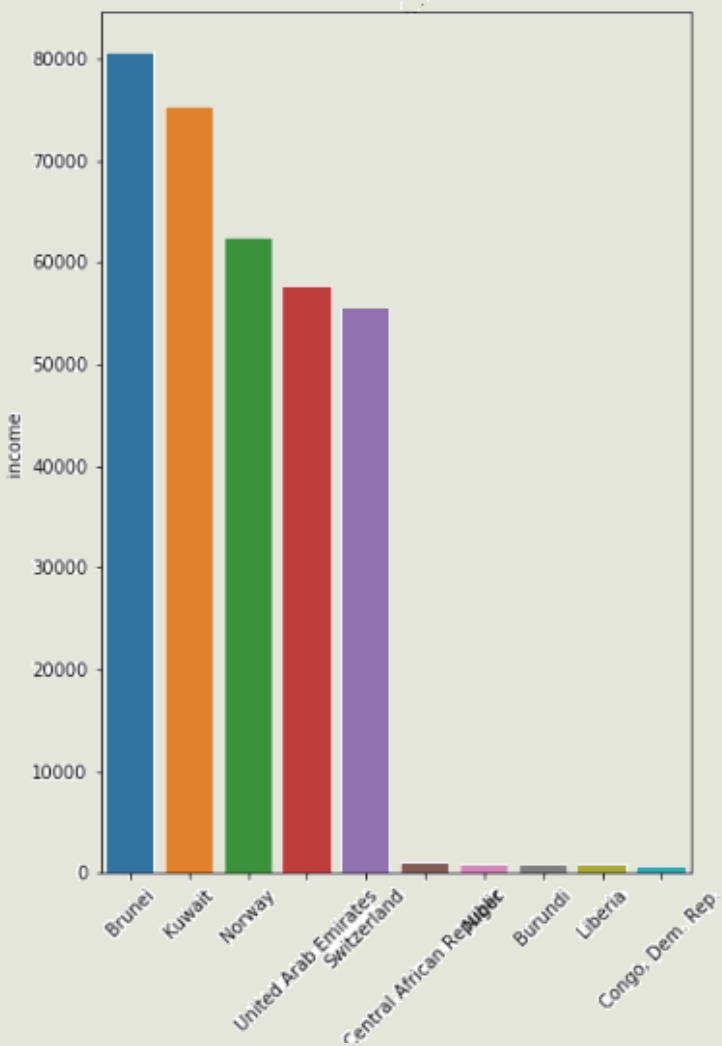

There are outliers in data for the following columns:

- **Imports**
- **GDP**
- **Inflation**
- **Exports**

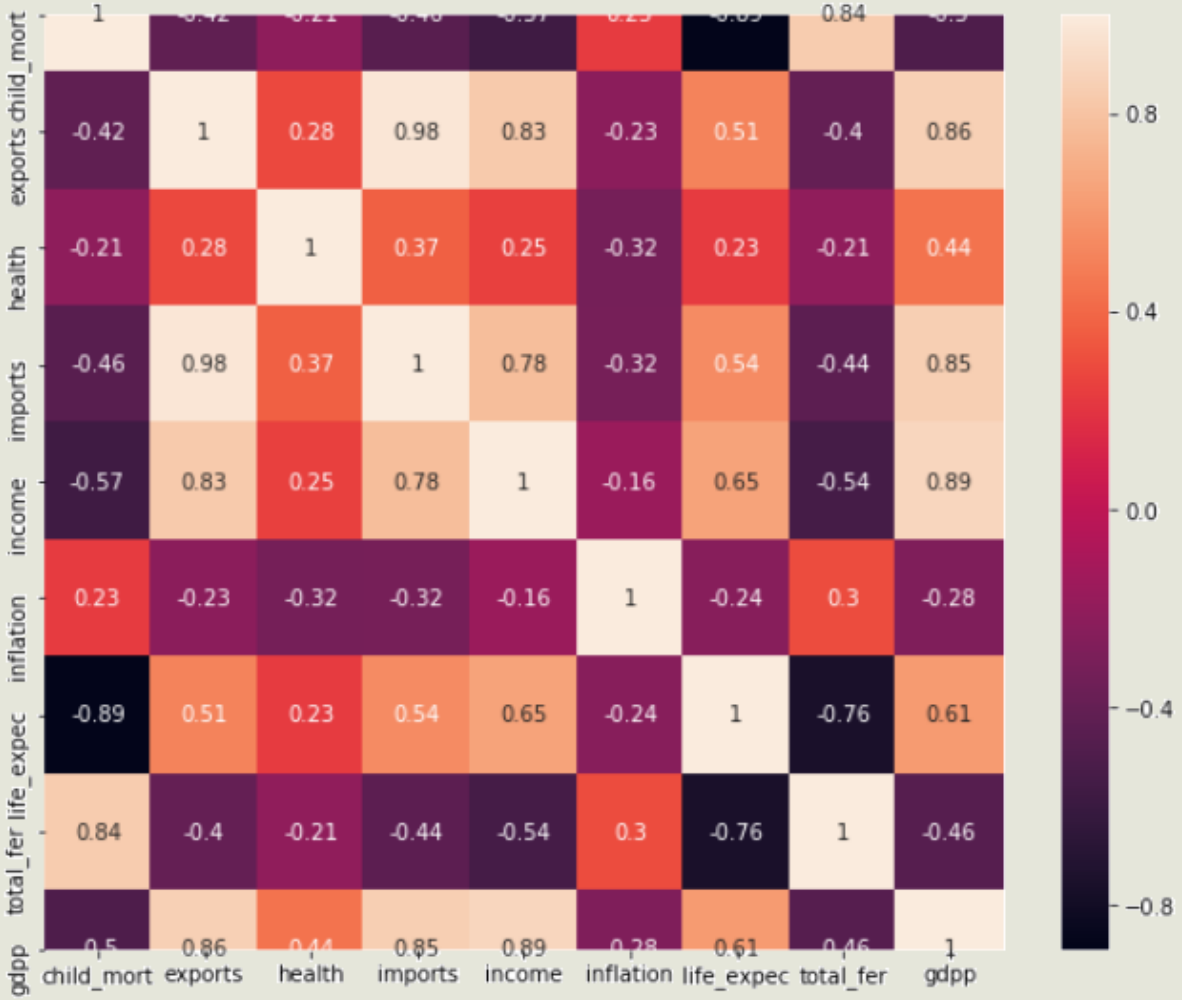
# TOP5 and Bottom5 countries for Child mortality, Income and GDP



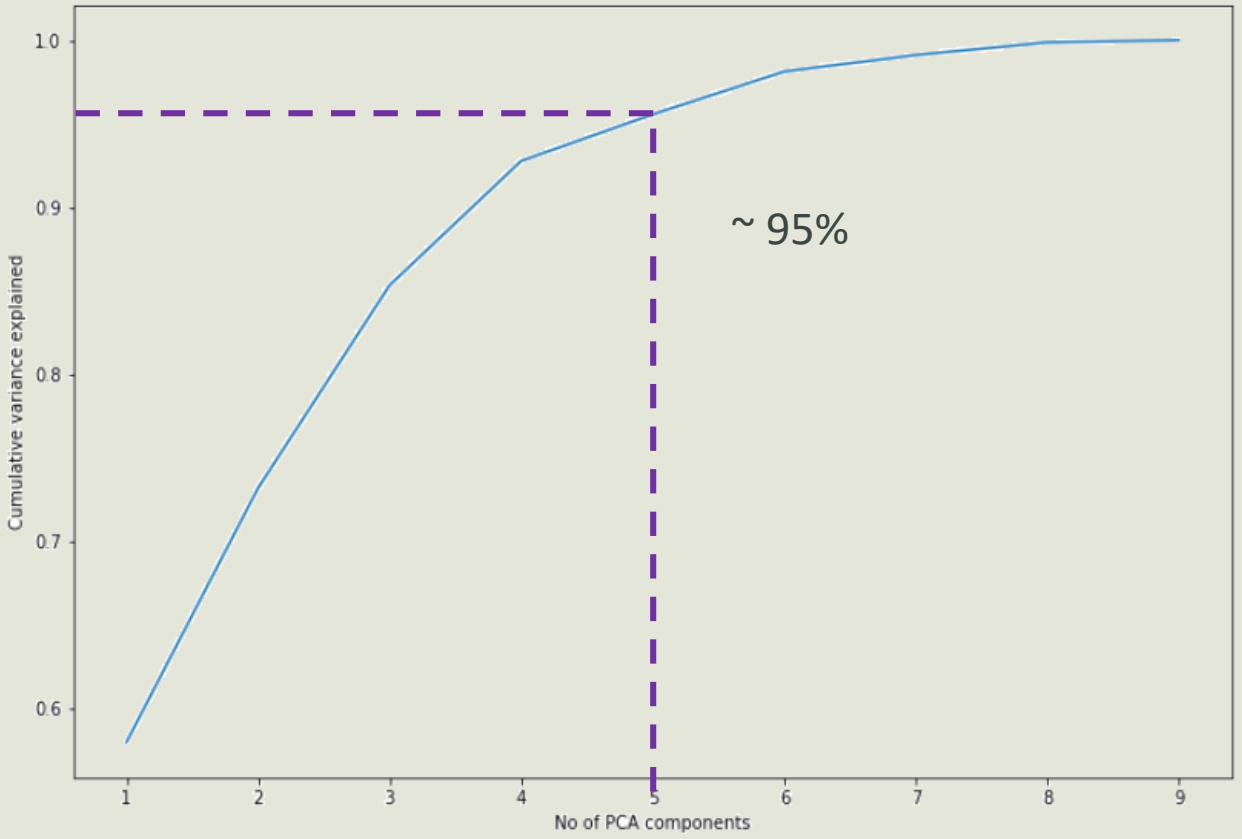



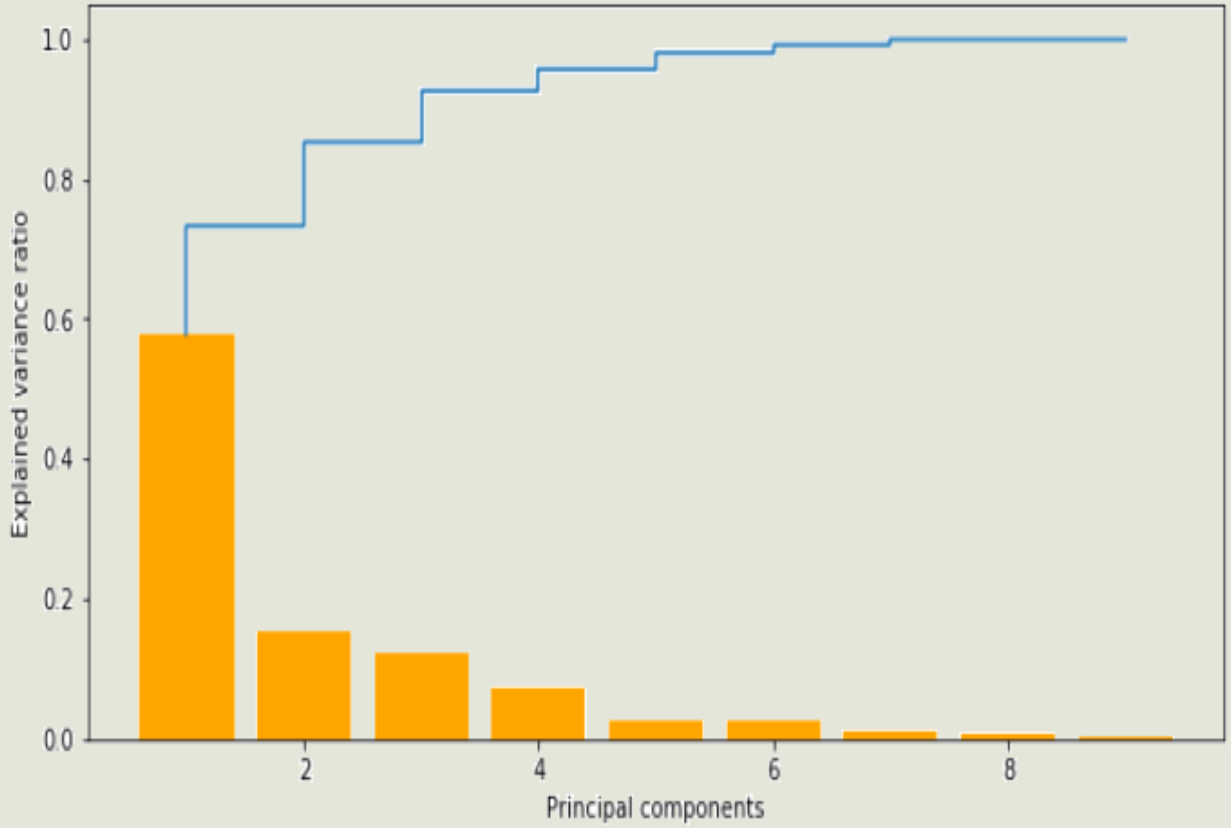
# Correlation between variables




- Import and Export have very high positive correlation with Income.
- Child Mortality and Life expectance are very high negative correlation.
- Income and GDP have very high positive correlation.
- Inflation has very less correlation with most of the parameters

# Scree Plot to find the optimum no of components for PCA

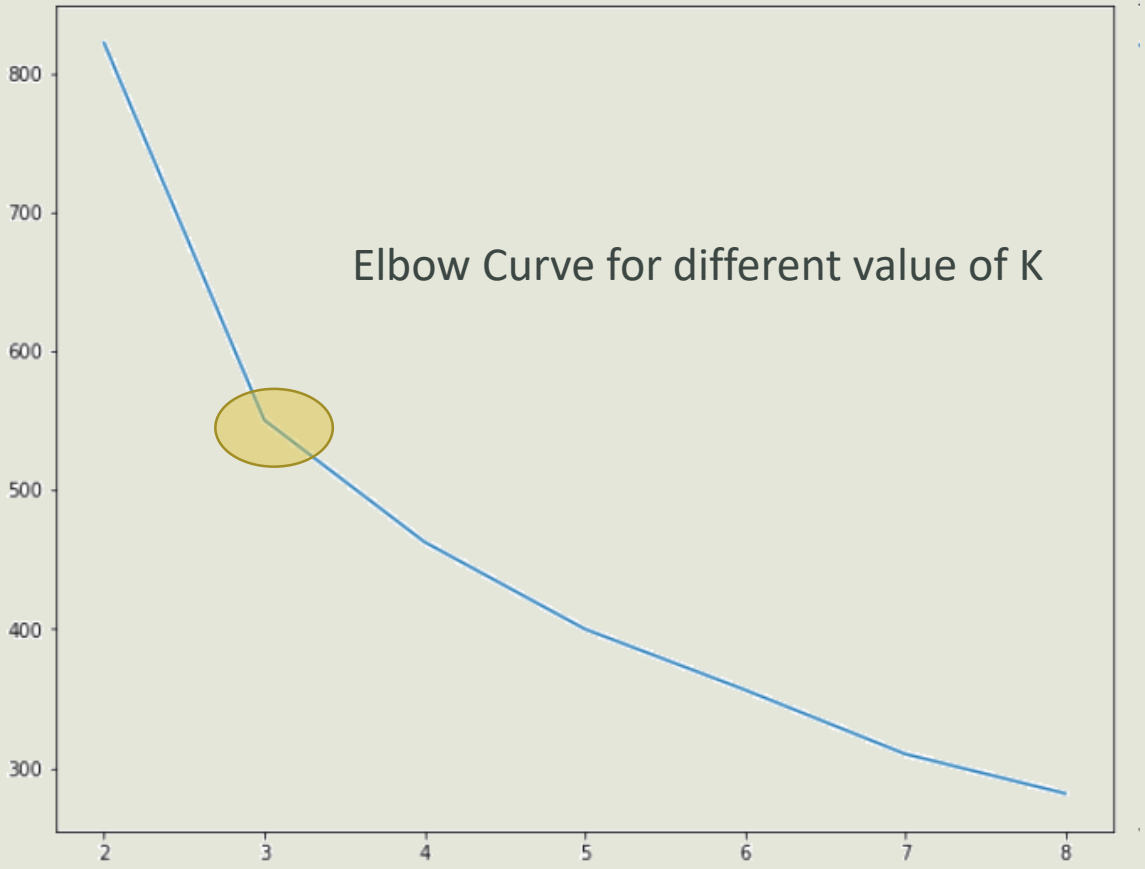


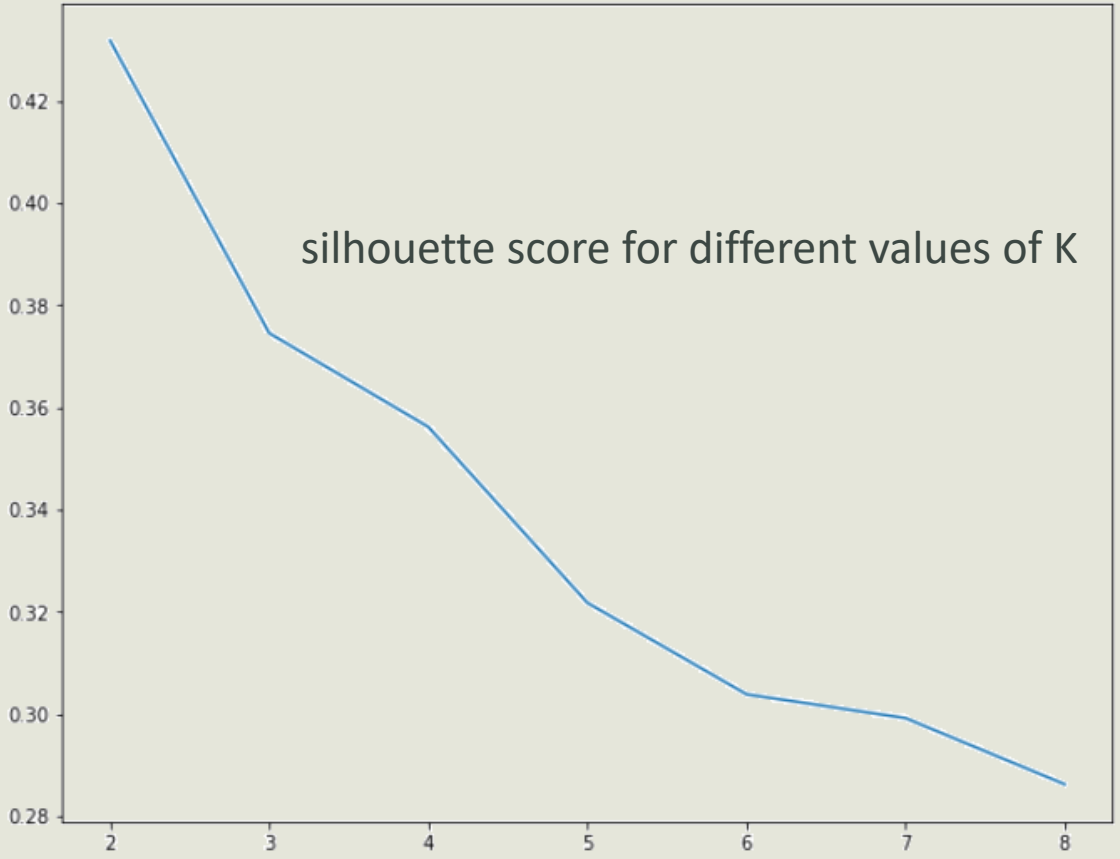




**Based on Scree plot 95% of the variability explained by 5 components**

**The no of complements selected for PCA is 5**

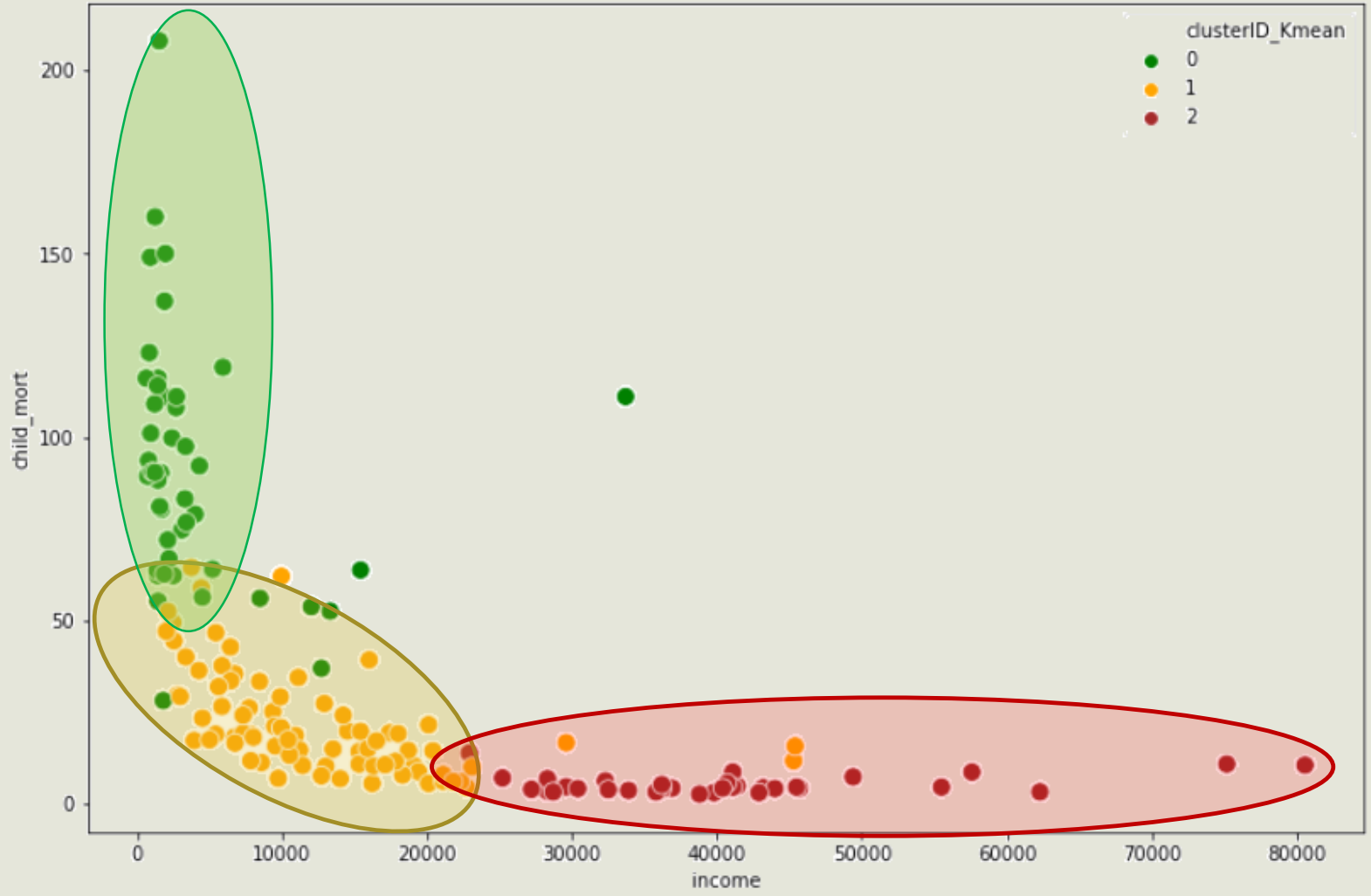
# Kmeans Clustering



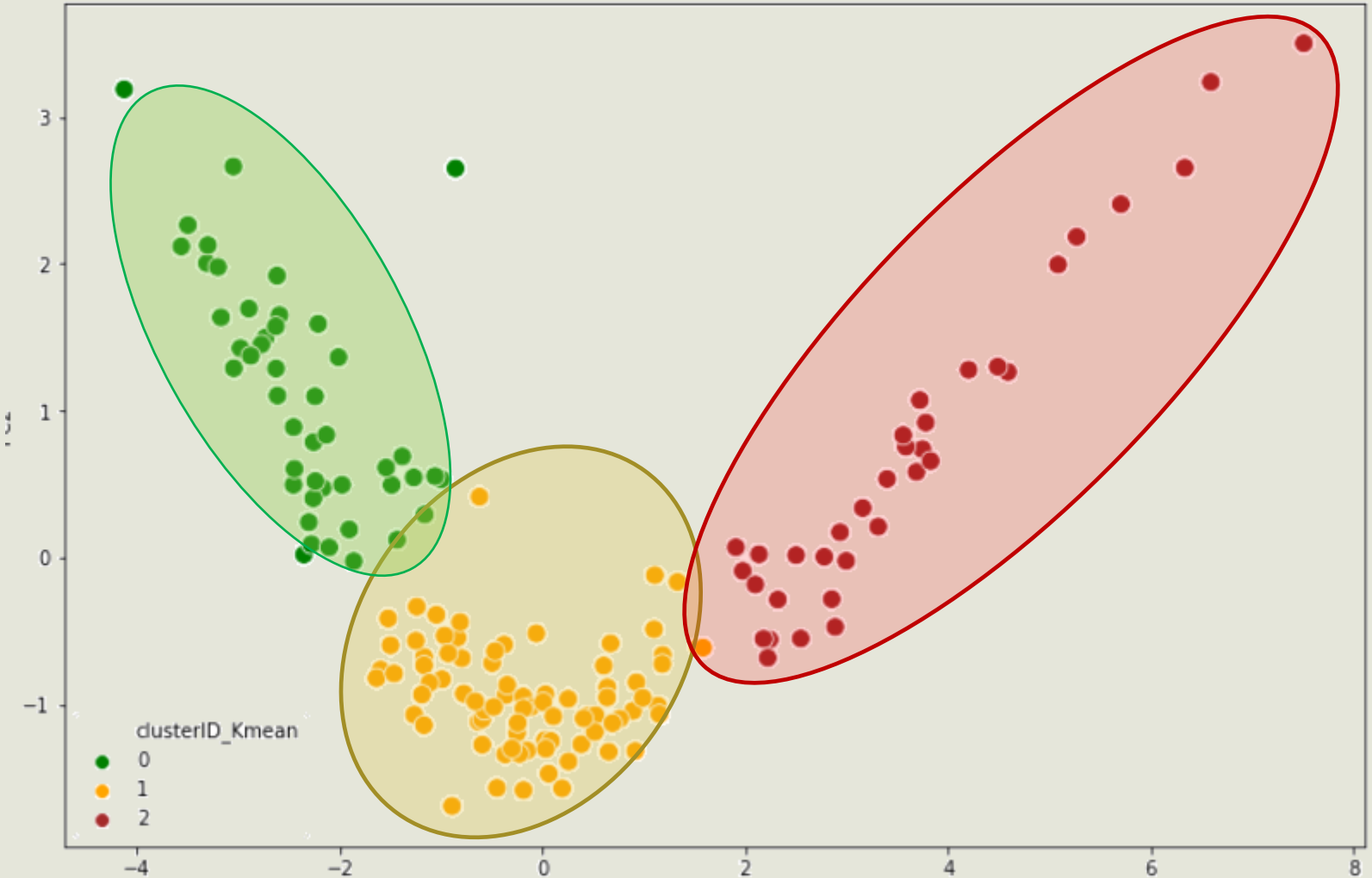




**The no of clusters (K) for Kmeans Clustering is considered as 3**

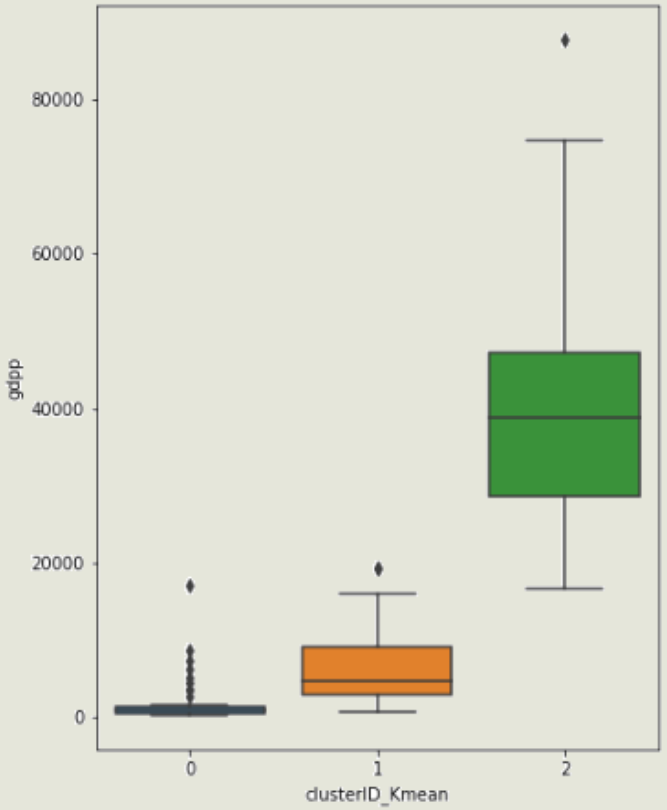
# Clusters created with Kmeans Clustering

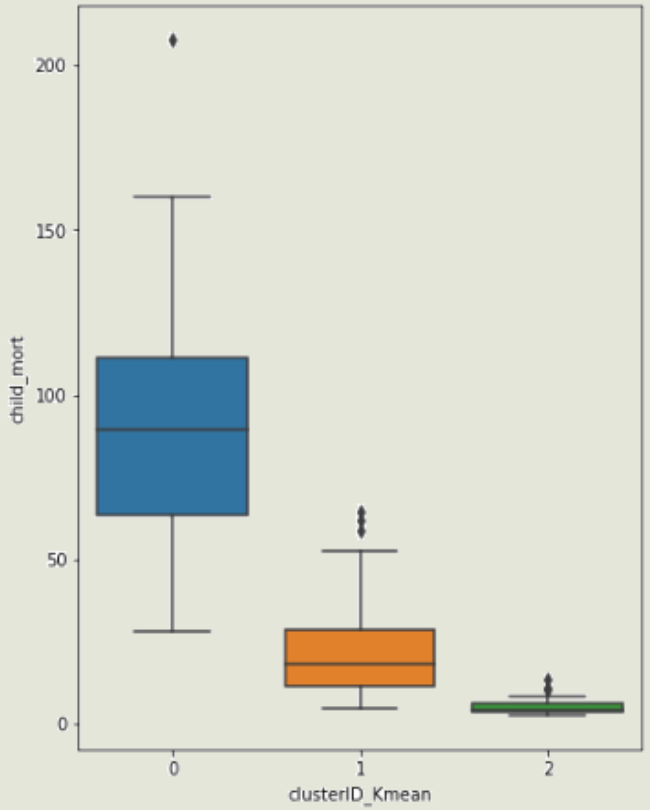


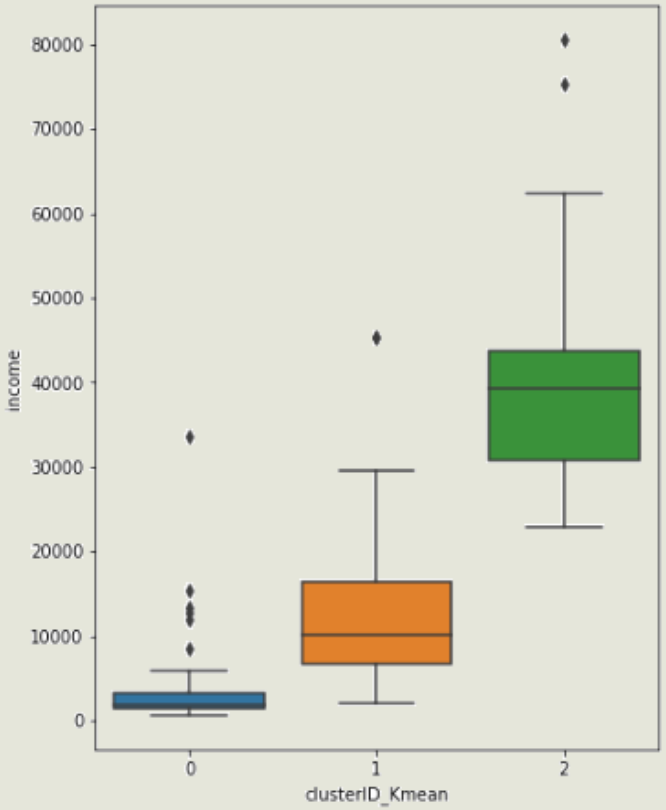
# Clusters created for PCA components - Kmeans



# Clusters



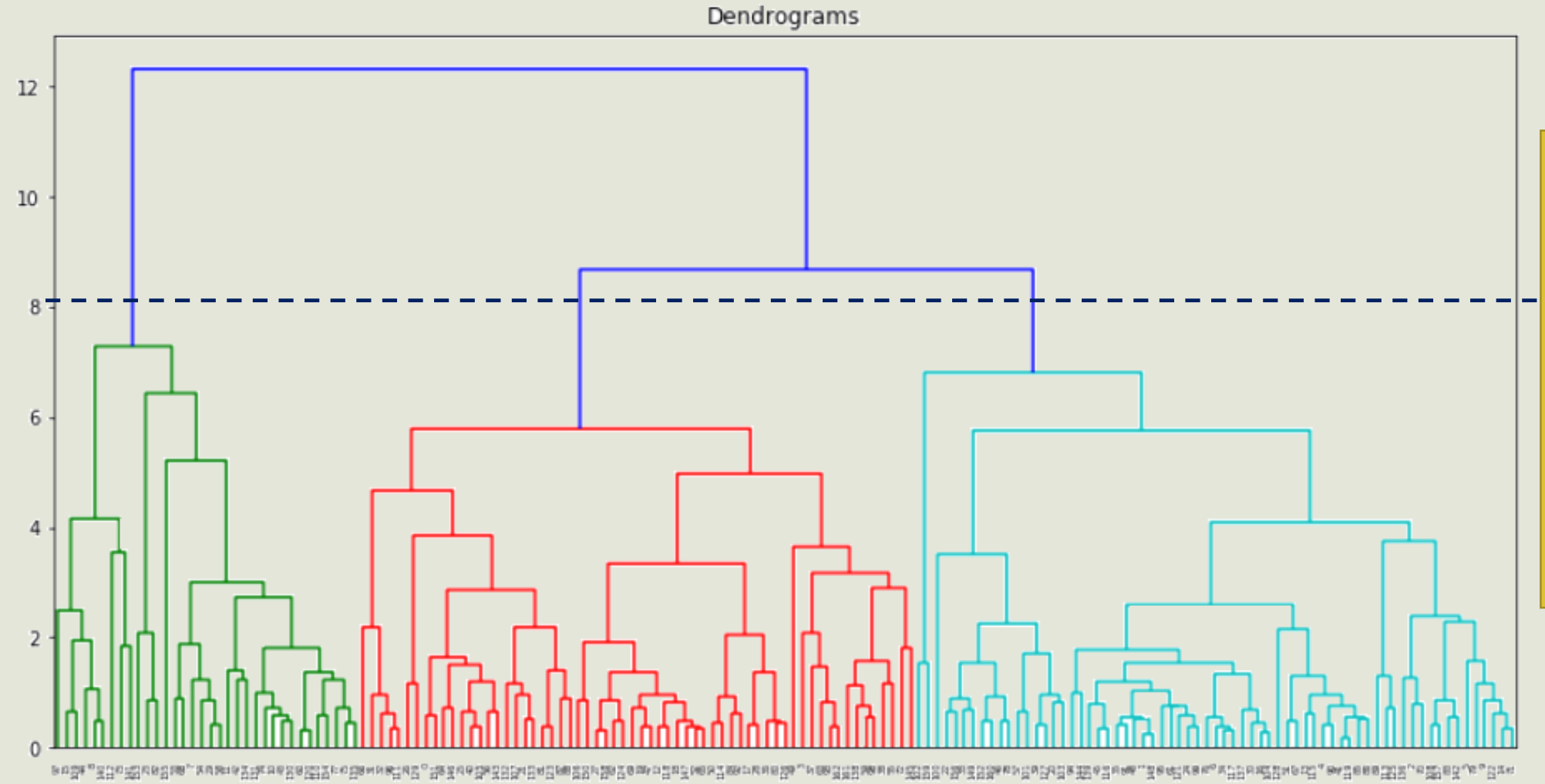

**Cluster 0: (Under Developed)**
Low GDP
Low Income
High Child mortality

**Cluster 2 : (Developed)**

High GDP
High Income
Low child Mortality

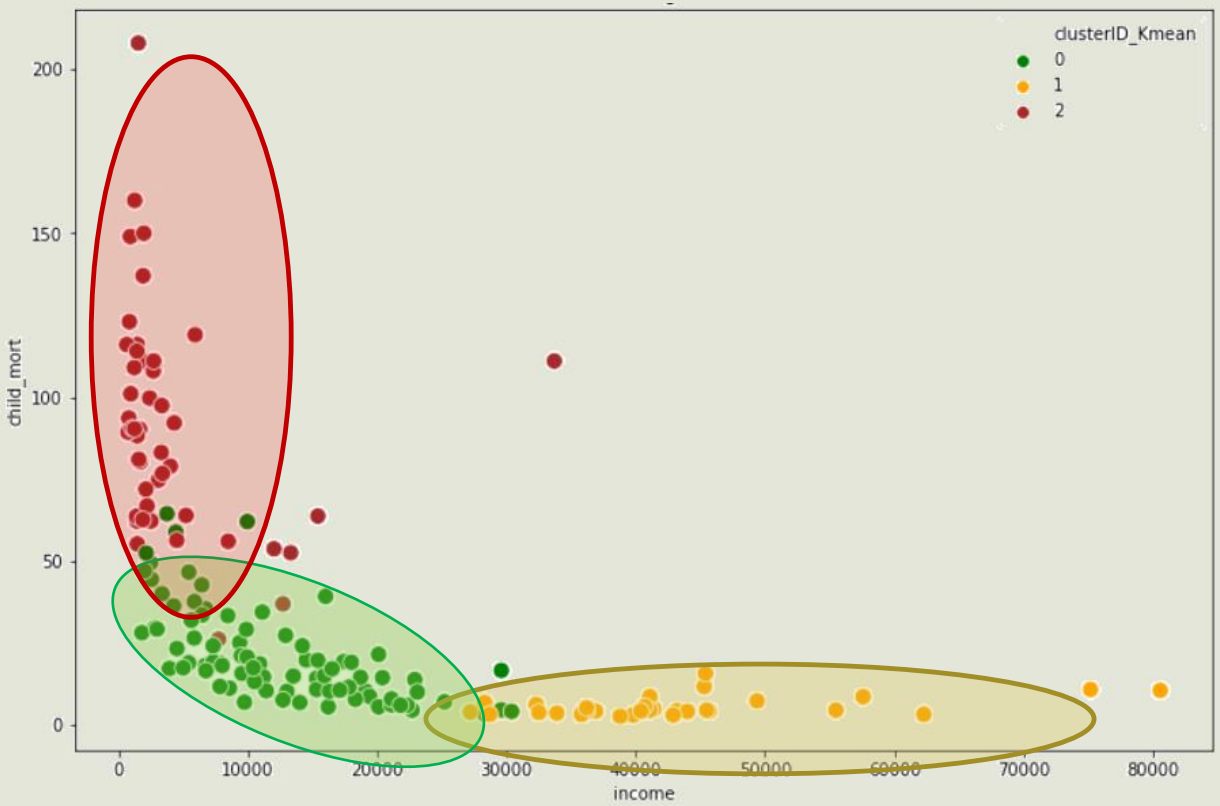
**Countries in Cluster 0 direst need of aid.**

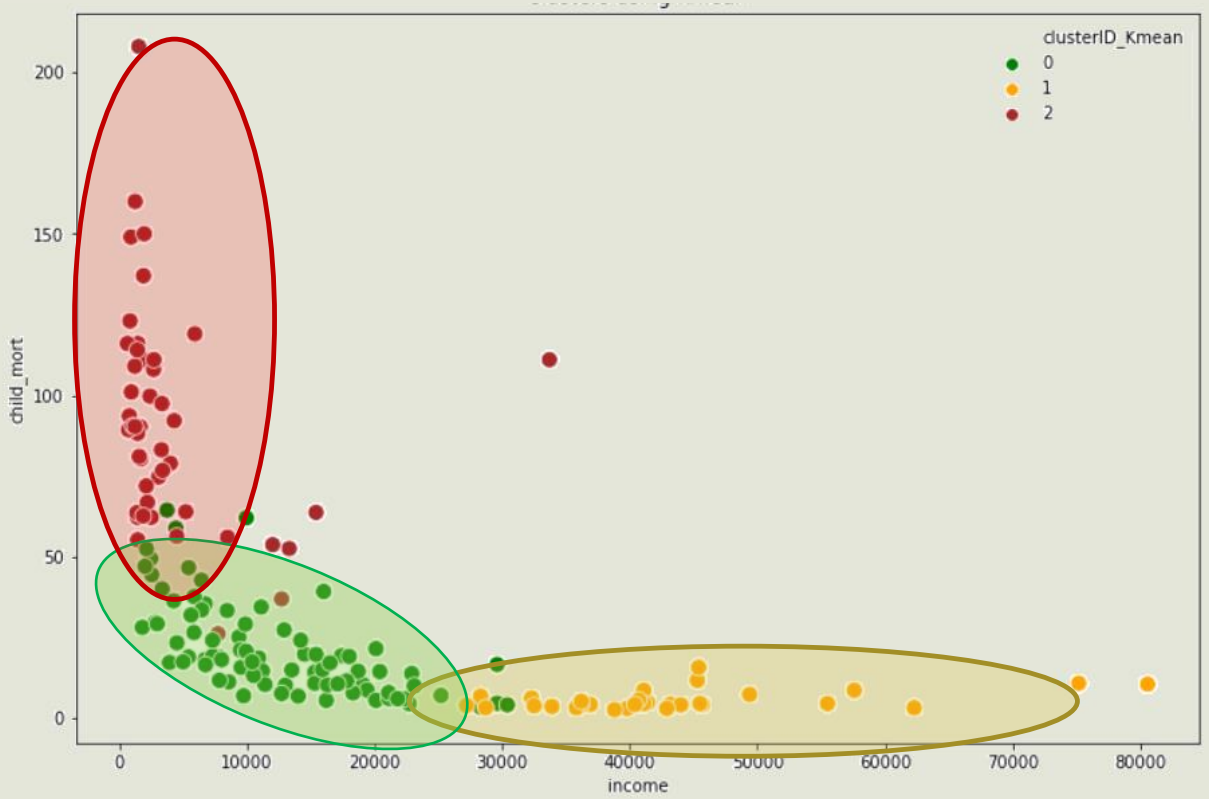
# Dendrograms for Hierarchical Clustering



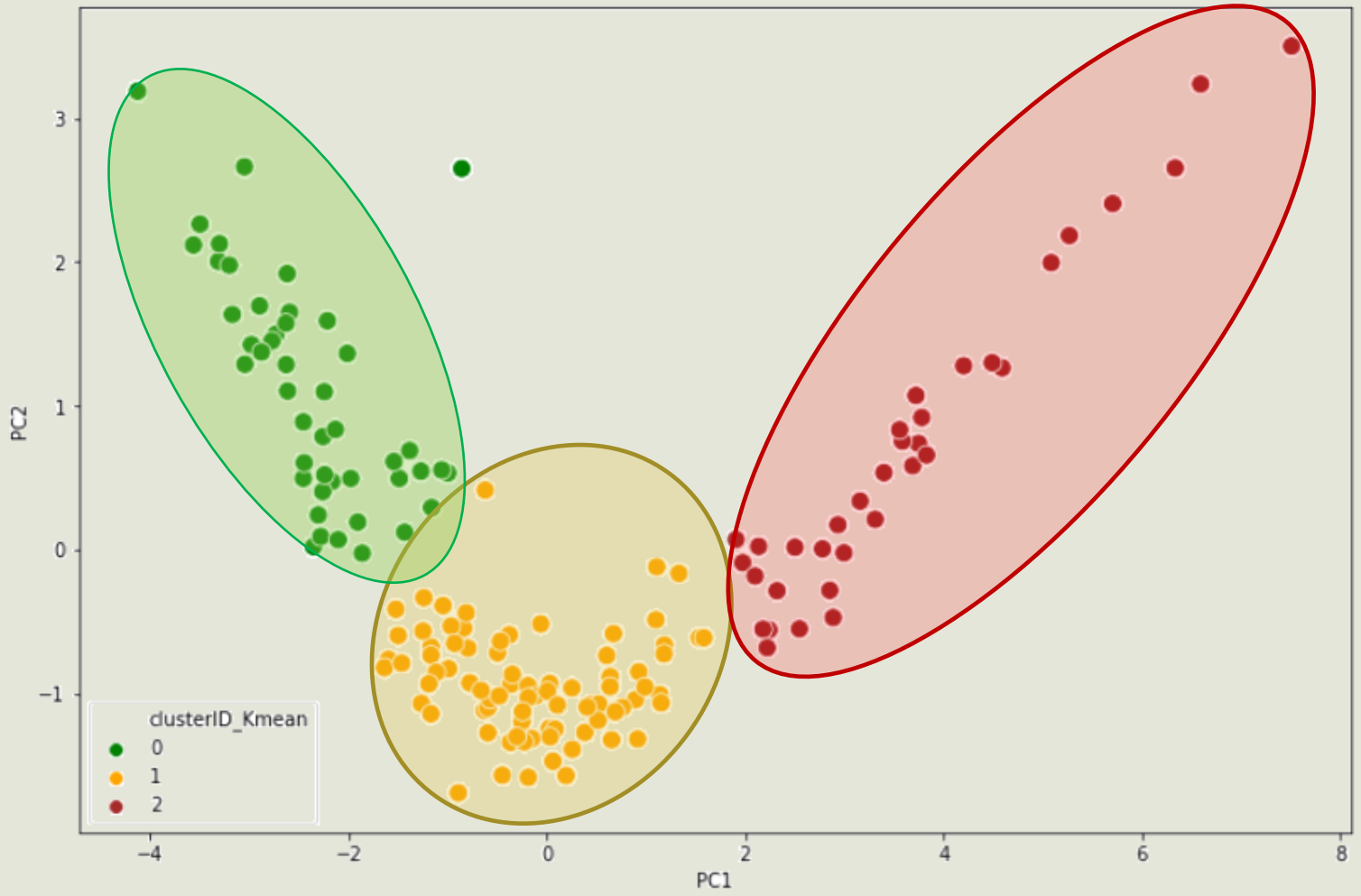

**“Complete” linkage type selected for the dendrogram as it gives distinct clusters**

# Clusters created with Hierarchical Clustering

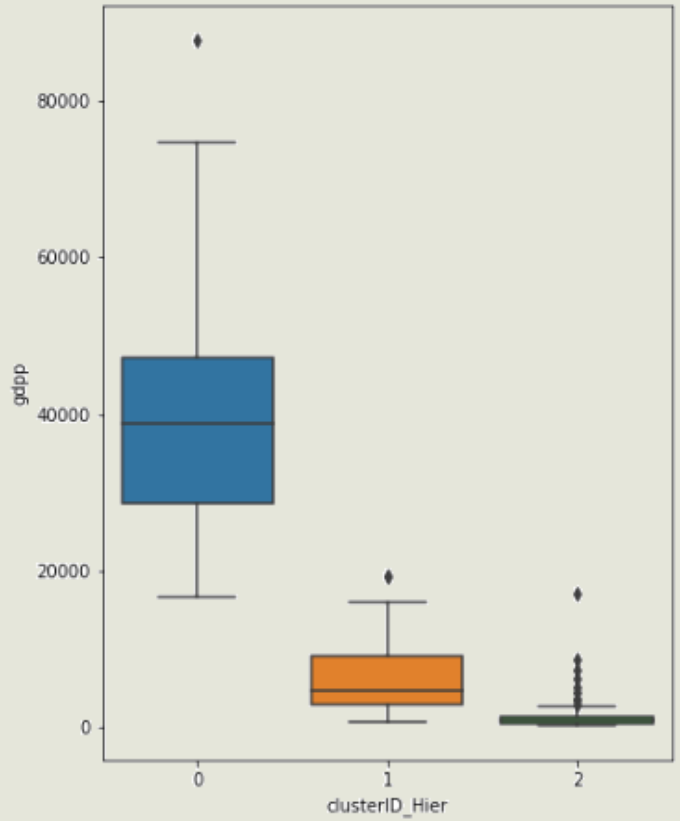


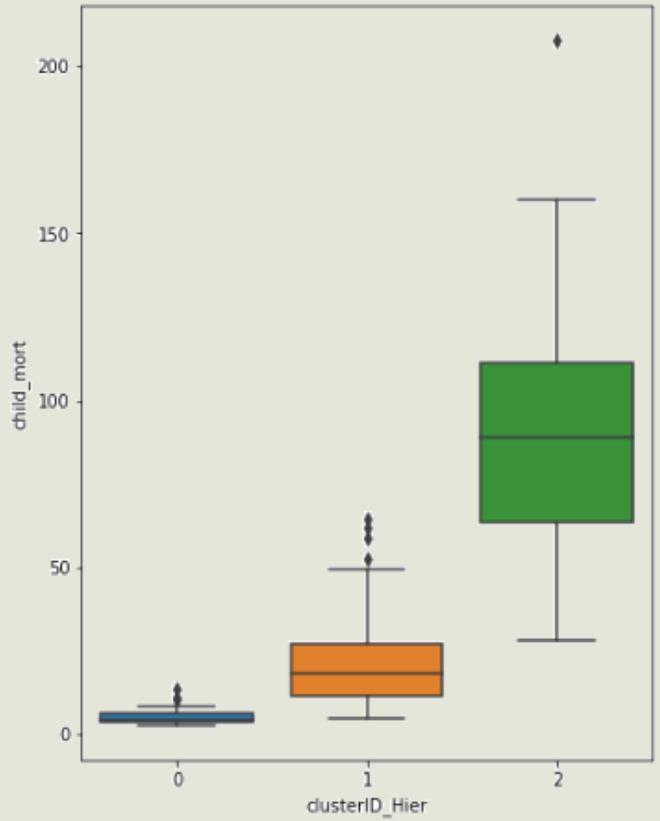



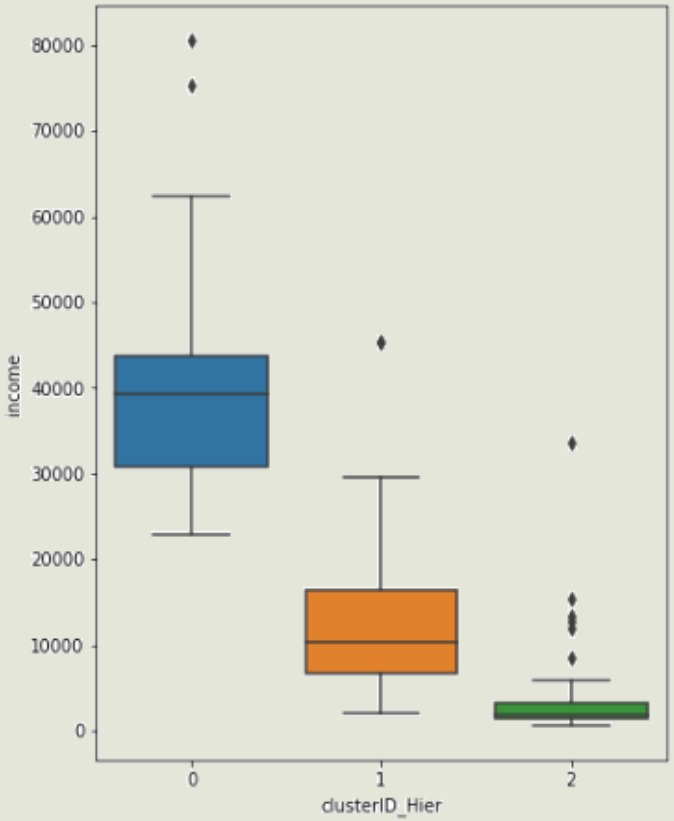
# Clusters created for PCA components



# Clusters




**<u>Cluster 2: (Under Developed)</u>**
Low GDP
Low Income
High Child mortality

**<u>Cluster 0 : (Developed)</u>**

High GDP
High Income
Low child Mortality

**Both Kmeans and Hierarchical Clustering methods provided similar clusters. Countries in Cluster 2 from Hierarchical method are direst need of aid.**

# Bottom 10 Counties with lower GDP

| country | child_mort | exports | health | imports | income | inflation | life_expec | total_fer | gdpp |
|---|---|---|---|---|---|---|---|---|---|
| Liberia | 89.3 | 62.46 | 11.80 | 302.80 | 700 | 5.47 | 60.8 | 5.02 | 327 |
| Congo, Dem. Rep. | 116.0 | 137.27 | 7.91 | 165.66 | 609 | 20.80 | 57.5 | 6.54 | 334 |
| Niger | 123.0 | 77.26 | 5.16 | 170.87 | 814 | 2.55 | 58.8 | 7.49 | 348 |
| Sierra Leone | 160.0 | 67.03 | 13.10 | 137.66 | 1220 | 17.20 | 55.0 | 5.20 | 399 |
| Madagascar | 62.2 | 103.25 | 3.77 | 177.59 | 1390 | 8.79 | 60.8 | 4.60 | 413 |
| Mozambique | 101.0 | 131.99 | 5.21 | 193.58 | 918 | 7.64 | 54.5 | 5.56 | 419 |
| Central African Republic | 149.0 | 52.63 | 3.98 | 118.19 | 888 | 2.01 | 47.5 | 5.21 | 446 |
| Malawi | 90.5 | 104.65 | 6.59 | 160.19 | 1030 | 12.10 | 53.1 | 5.31 | 459 |
| Eritrea | 55.2 | 23.09 | 2.66 | 112.31 | 1420 | 11.60 | 61.7 | 4.61 | 482 |

# TOP 5 Countries with High Child Mortality and Lowest GDP

| country | child_mort | exports | health | imports | income | inflation | life_expec | total_fer | gdpp |
|---|---|---|---|---|---|---|---|---|---|
| Sierra Leone | 160.0 | 67.03 | 13.10 | 137.66 | 1220 | 17.20 | 55.0 | 5.20 | 399 |
| Central African Republic | 149.0 | 52.63 | 3.98 | 118.19 | 888 | 2.01 | 47.5 | 5.21 | 446 |
| Niger | 123.0 | 77.26 | 5.16 | 170.87 | 814 | 2.55 | 58.8 | 7.49 | 348 |
| Congo, Dem. Rep. | 116.0 | 137.27 | 7.91 | 165.66 | 609 | 20.80 | 57.5 | 6.54 | 334 |
| Mozambique | 101.0 | 131.99 | 5.21 | 193.58 | 918 | 7.64 | 54.5 | 5.56 | 419 |

**As Income and GDP are highly co-related only GDP and Child Mortality were considered for selecting 5 countries.**

# Conclusion

- Based on the Clustering, Countries can be classified into 3 categories:
  - **Developed Countries (Cluster 0)**
  - **Middle income country (Cluster 1)**
  - **less economically developed country ( Cluster 2)**

- Top 5 Countries which is in need of AID
  - **Sierra Leone**
  - **Central African Republic**
  - **Niger**
  - **Congo, Dem. Rep.**
  - **Mozambique**