NGO HELP
CLUSTERING ASSIGNMENT

SAKSHI A. MANCHALWAR

CONTENTS

- Background
- Problem Statement
- Design Solution Approach
- Data Cleaning
  - ✓ Imputing Missing values
- K-Means Clustering
- Heirarchical Clustering
- Conclusion

Background

- HELP 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 with advocacy drives for funding purposes and raised around $ 10 million.
- Now CEO of NGO wants to choosing the countries that are in the direst need of aid.

Problem Statement

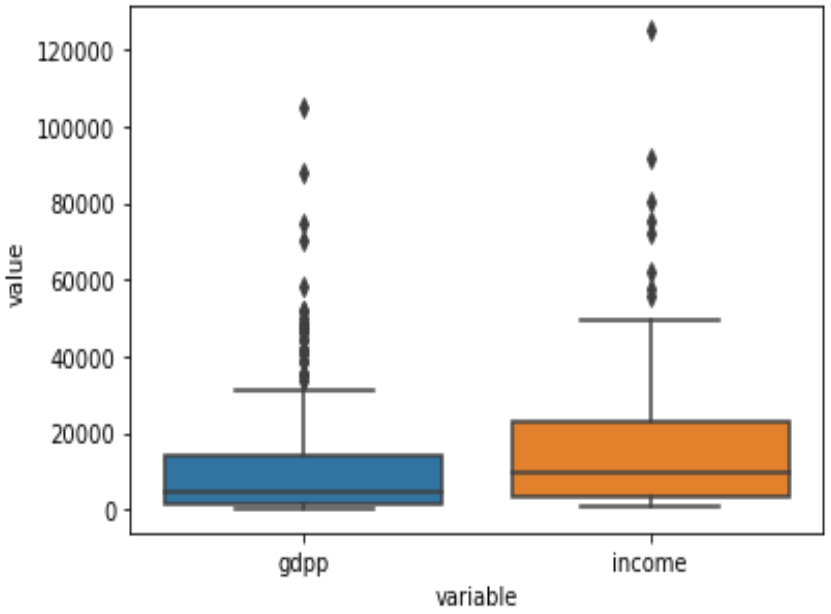
- Objective is to categorizes the countries using some socio-economic and health factors that determine the overall development of the country.
- Develop analytics to solutions classify countries which are in direst need of aid.
- Suggest the countries which the CEO needs to focus on the most.

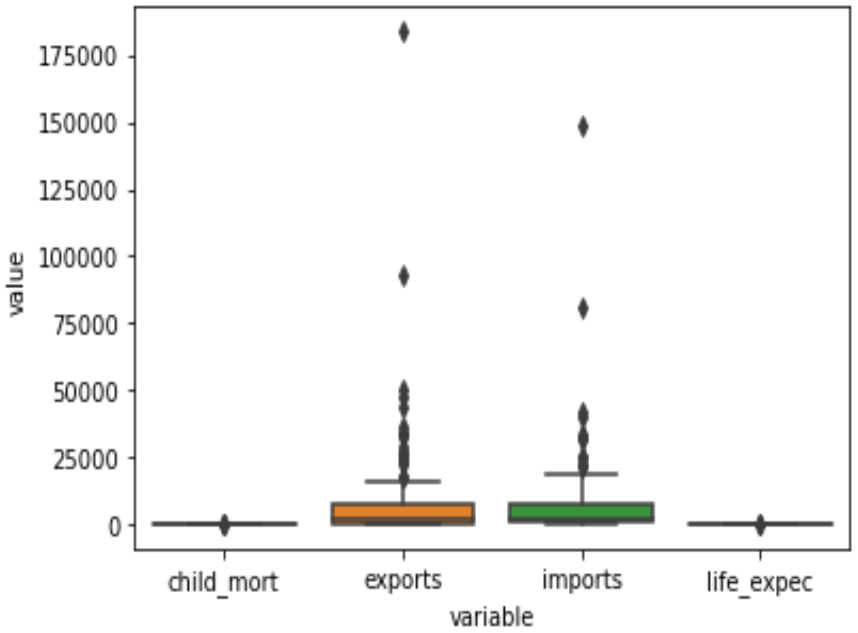
Solution Approach

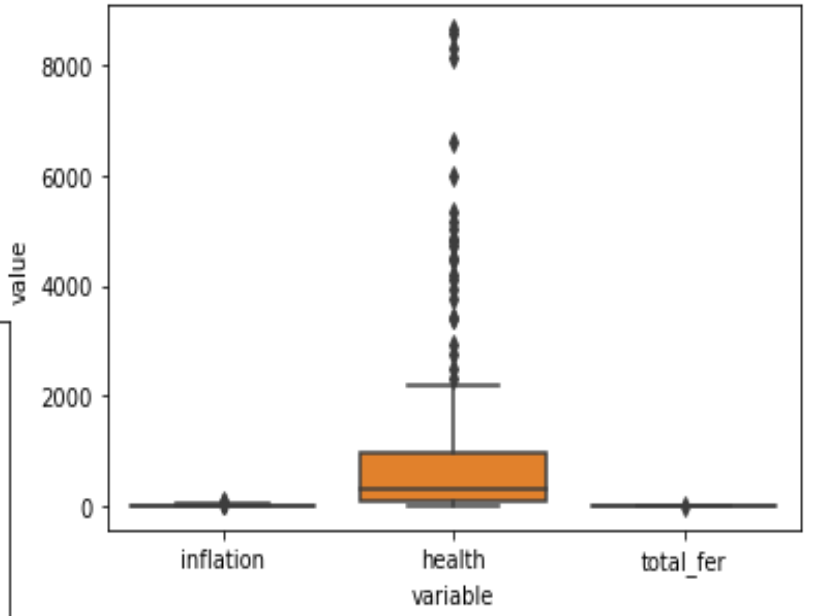
- Step 1: Reading and Understanding the Data
- Step 2: Data Cleansing • Missing Value check
- Step 3: Data Visualization
- Step 4: Data Preparation • Rescaling
- Step 5: Hopkins Statistics Test • Hopkins Score Calculation
- Step 6: Model Building • K-means Clustering • Elbow Curve • Silhouette Analysis • Hierarchical Clustering
- Step 8: Final Analysis • Final Country list Preparation

EDA

Outliers Treatment

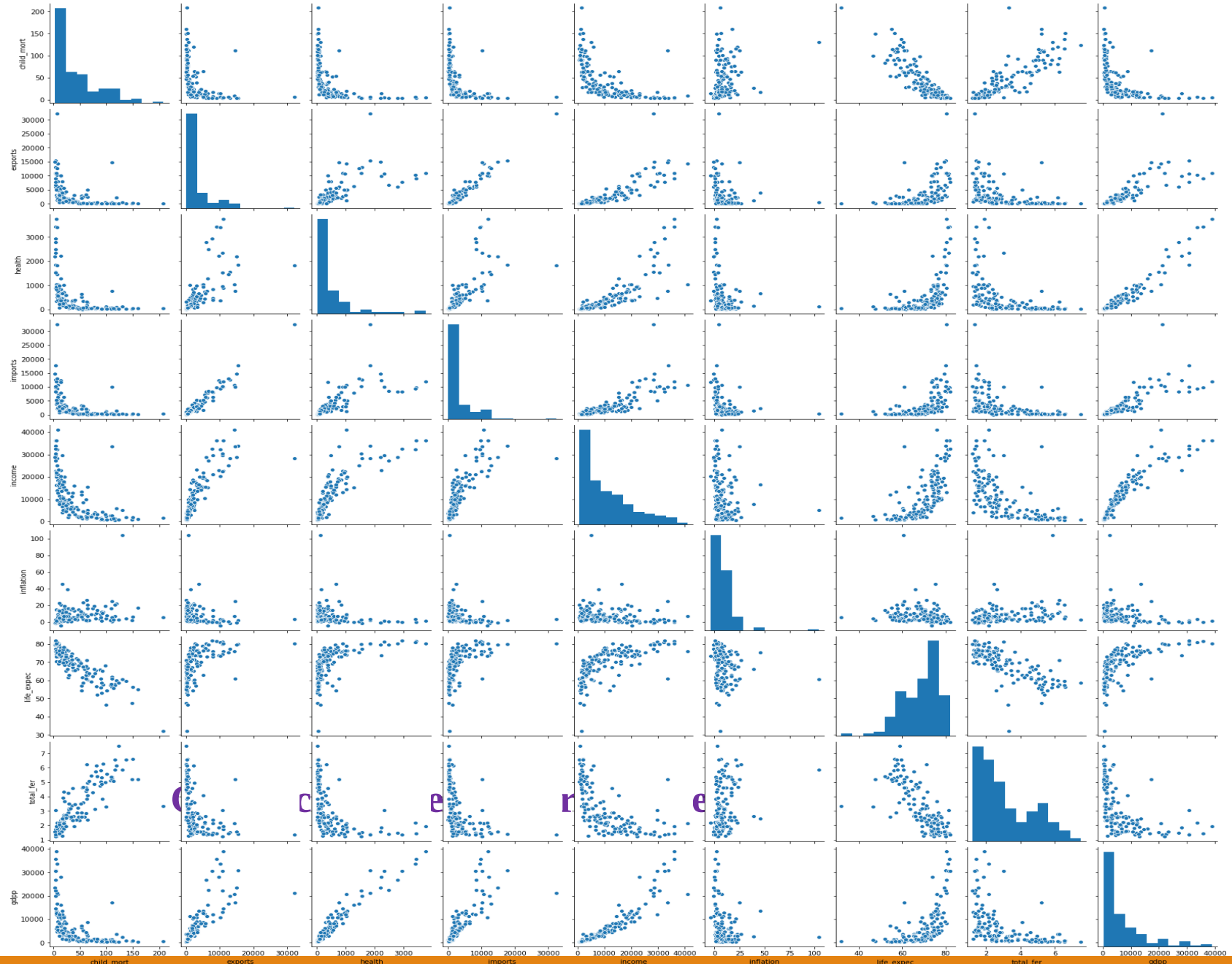


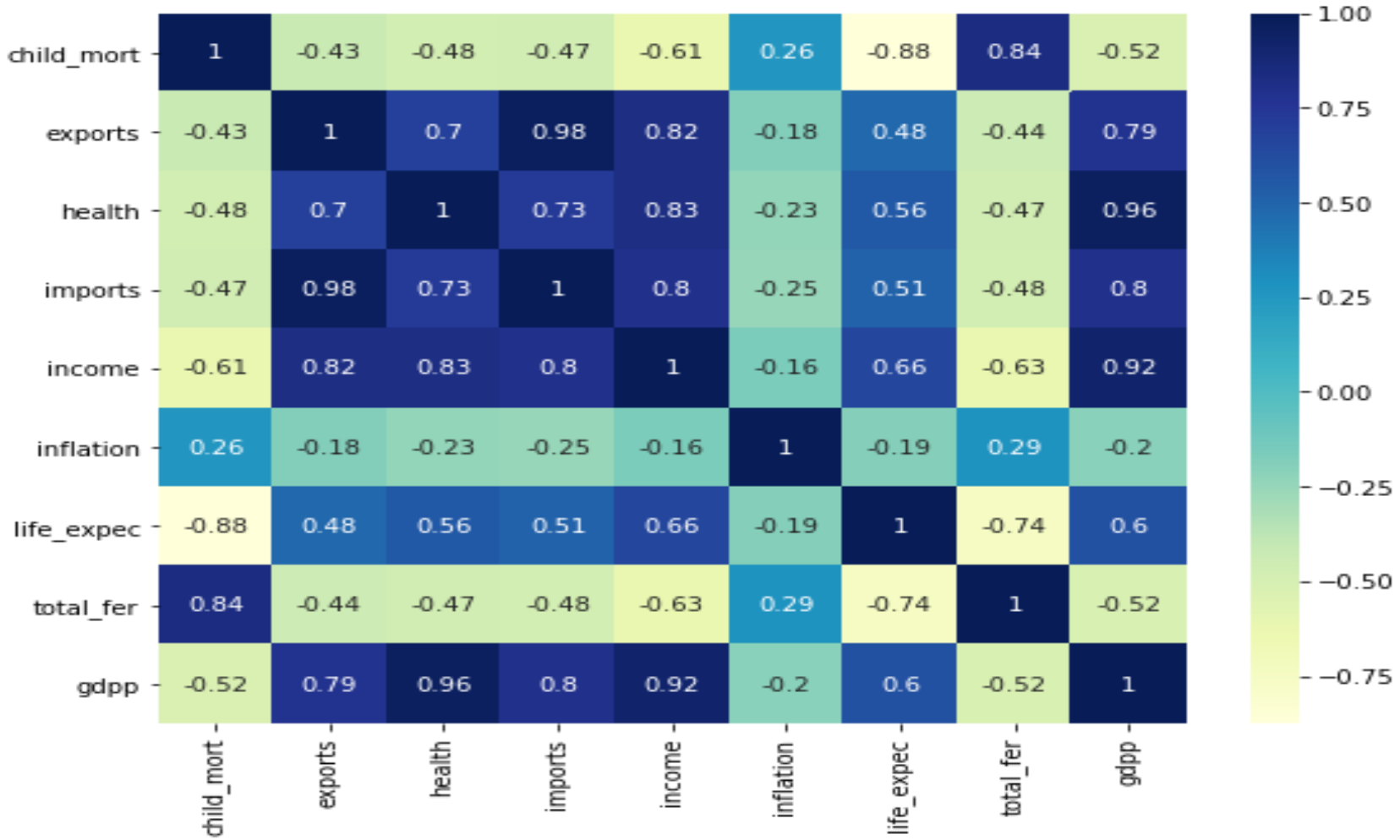




Removed Outliers from gdpp, income, child_mort using caping for lower and upper limit if we tried to remove more outliers we may loose some data which may be useful for business

Pair Plot showing relation between all numerical variables

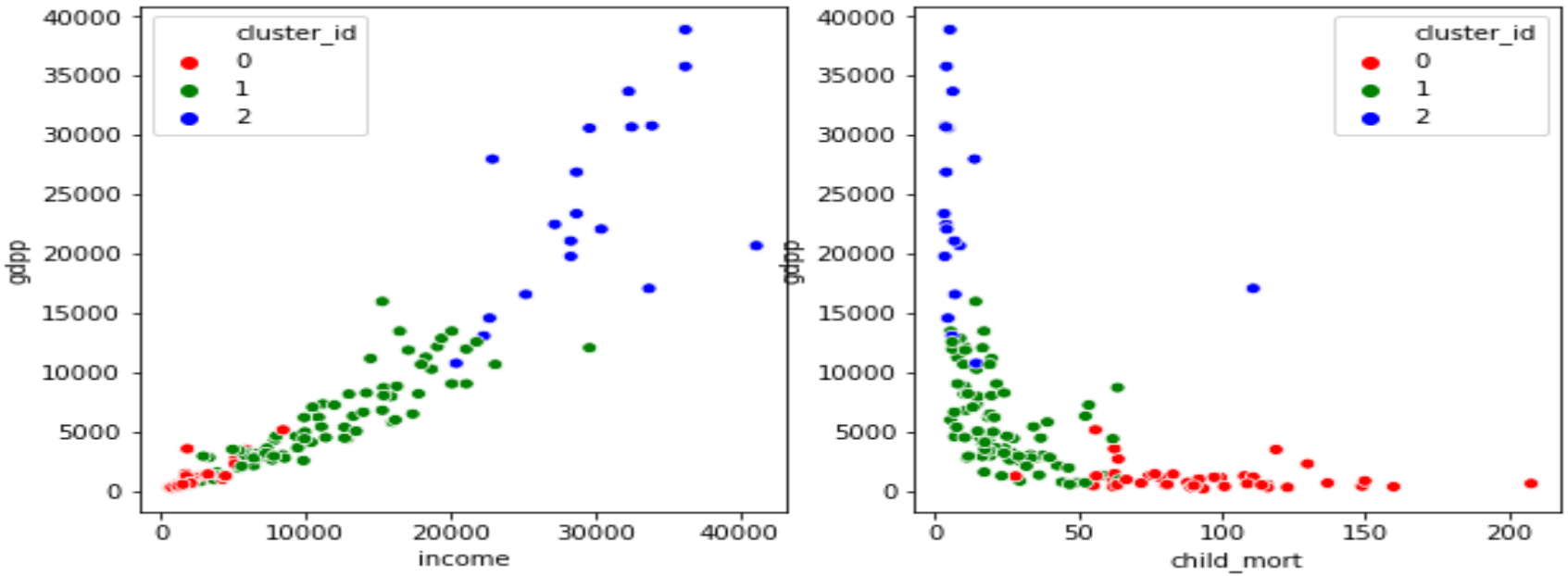




From above chart it is concluded that there is high positive correlation between total_fer and child_mort, between gdpp and income, and between imports and exports. strong negative correlation in life_expec and child_mort

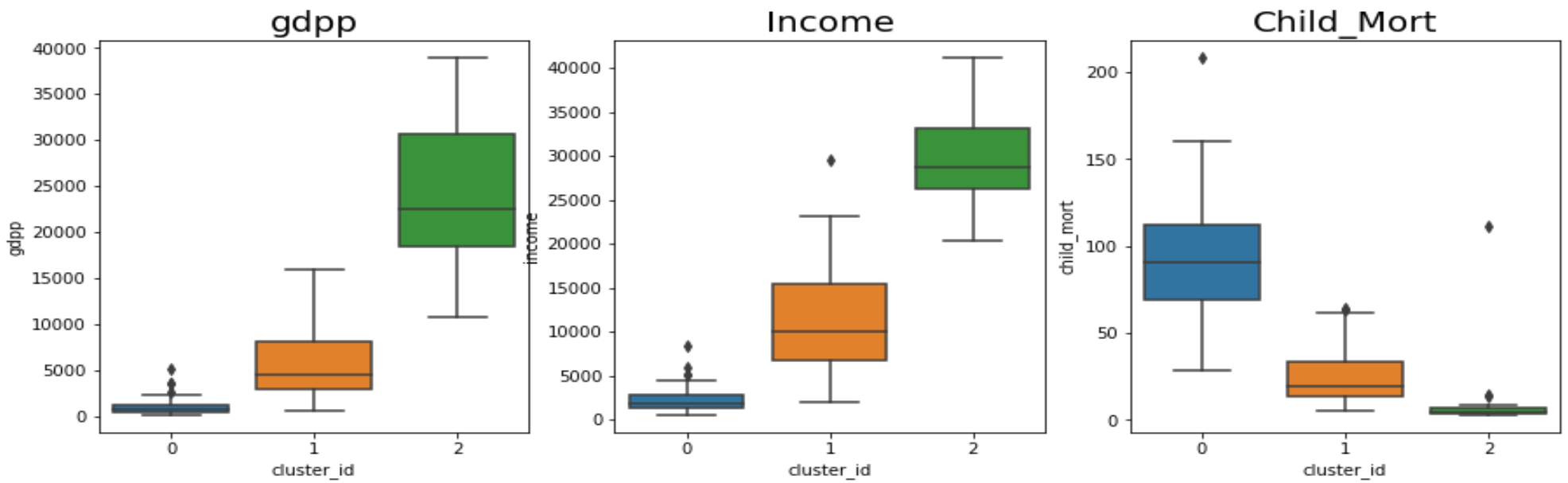
After performing Outlier Treatment Following Analysis Steps are performed:

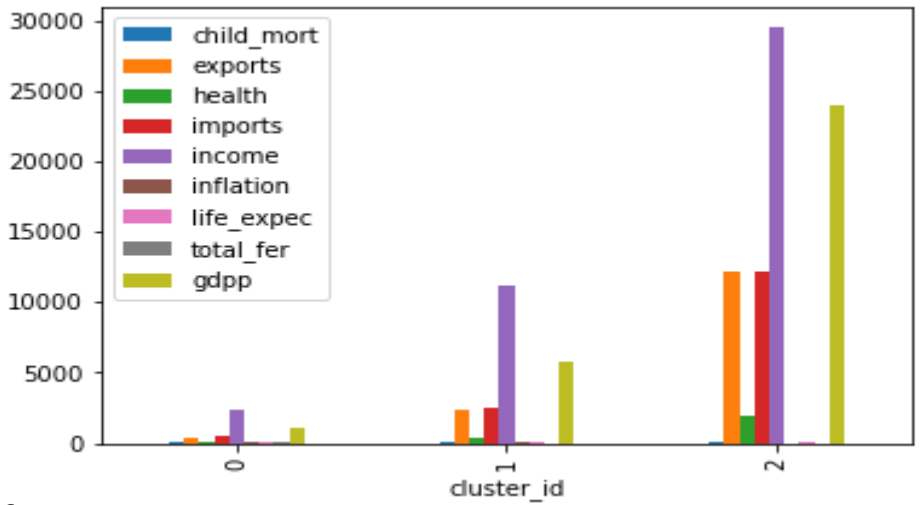
- Scaling Data: Using Standard Scaler
- Hopkins Score: Using Hopkins Statistics
- Optimal Number of Clusters: Using elbow method and silhouette score and taking 3 clusters for K-Means clusterring

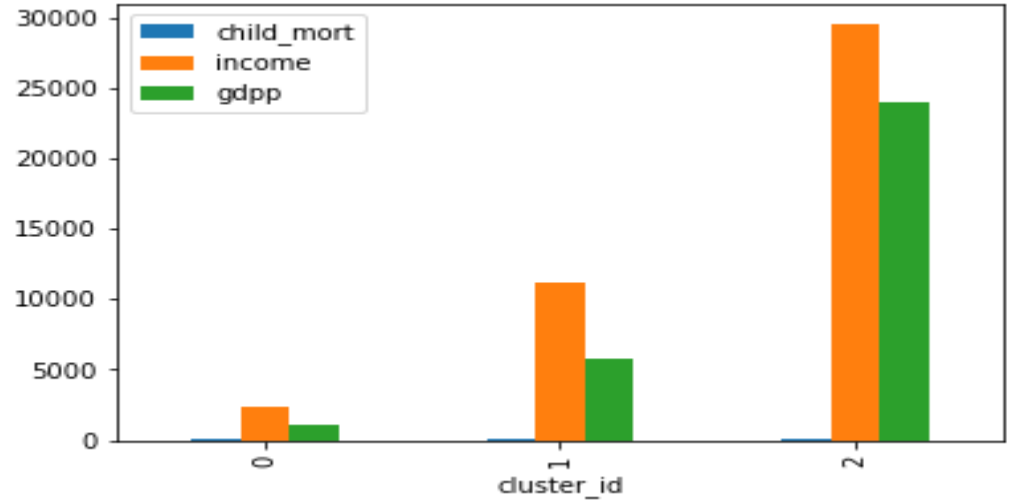


Clustering Pattern in K-Means Clusterring

Cluster Profiling







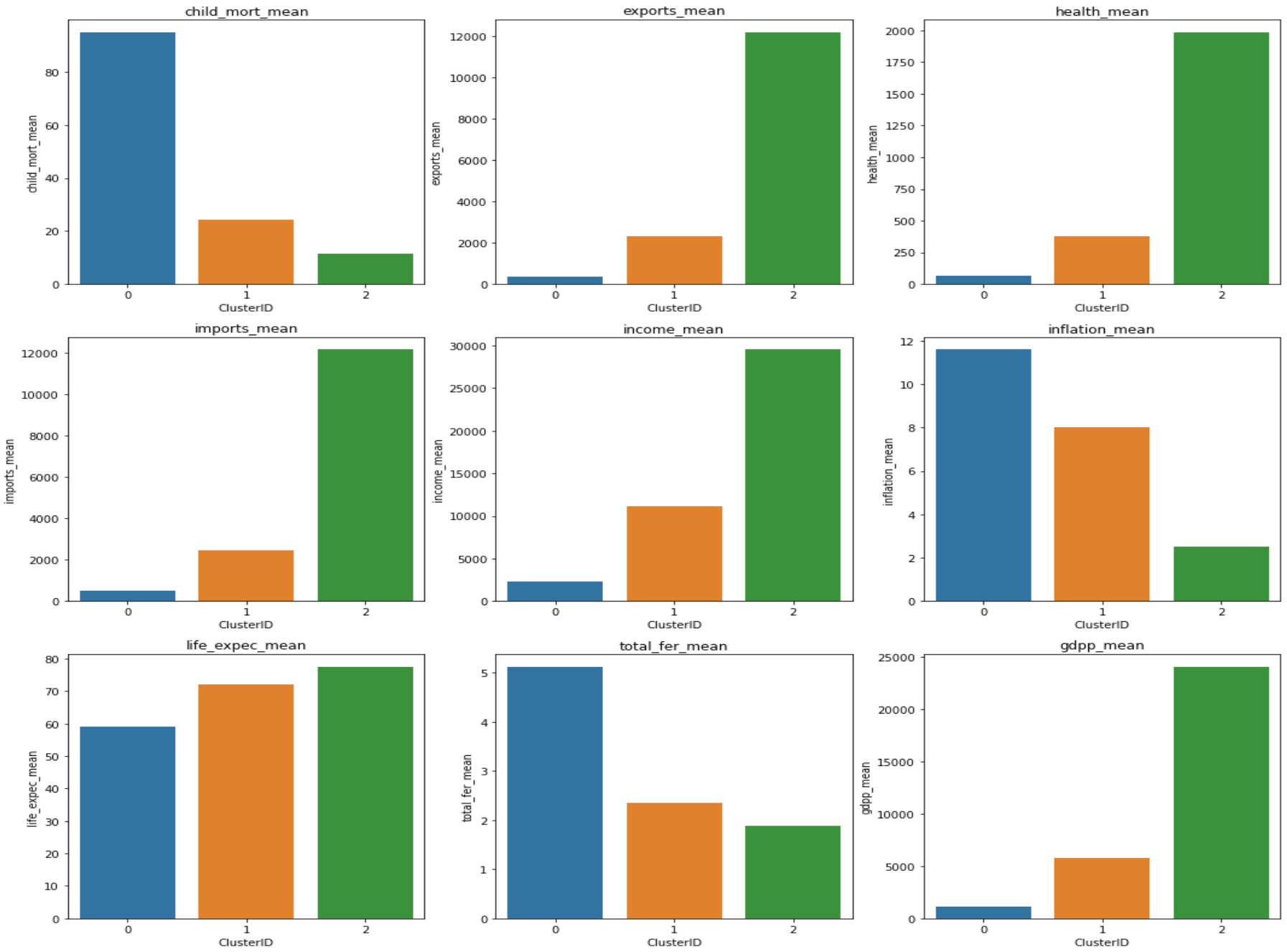
Cluster label 0 : Most of the countries in this boxplot are having little high gdpp and moderate income and child mortality

Cluster label 1 : Having highest gdpp and income with very low child mortality than all other cluster labels with some outliers

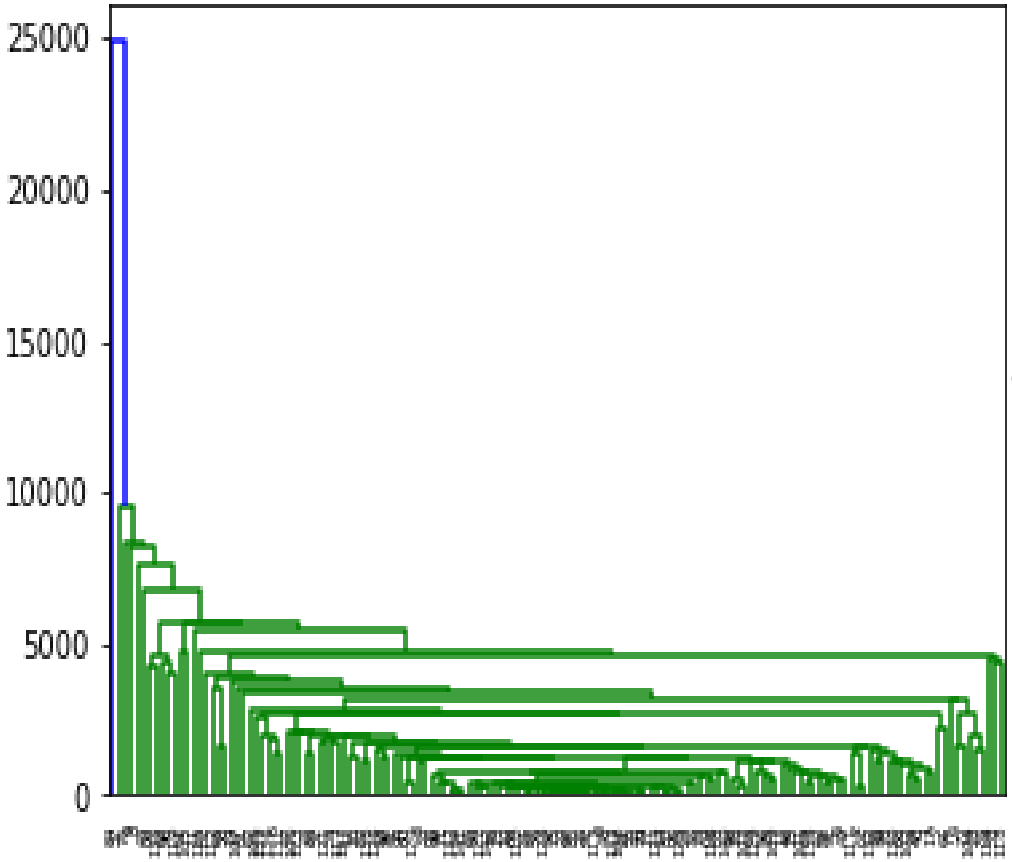
Clabel 2 : Most of the countries in this boxplot are having lowest gdpp and income with highest child mortality.

Looking at the graph we are certain that cluster 2 is cluster of concern. Because:

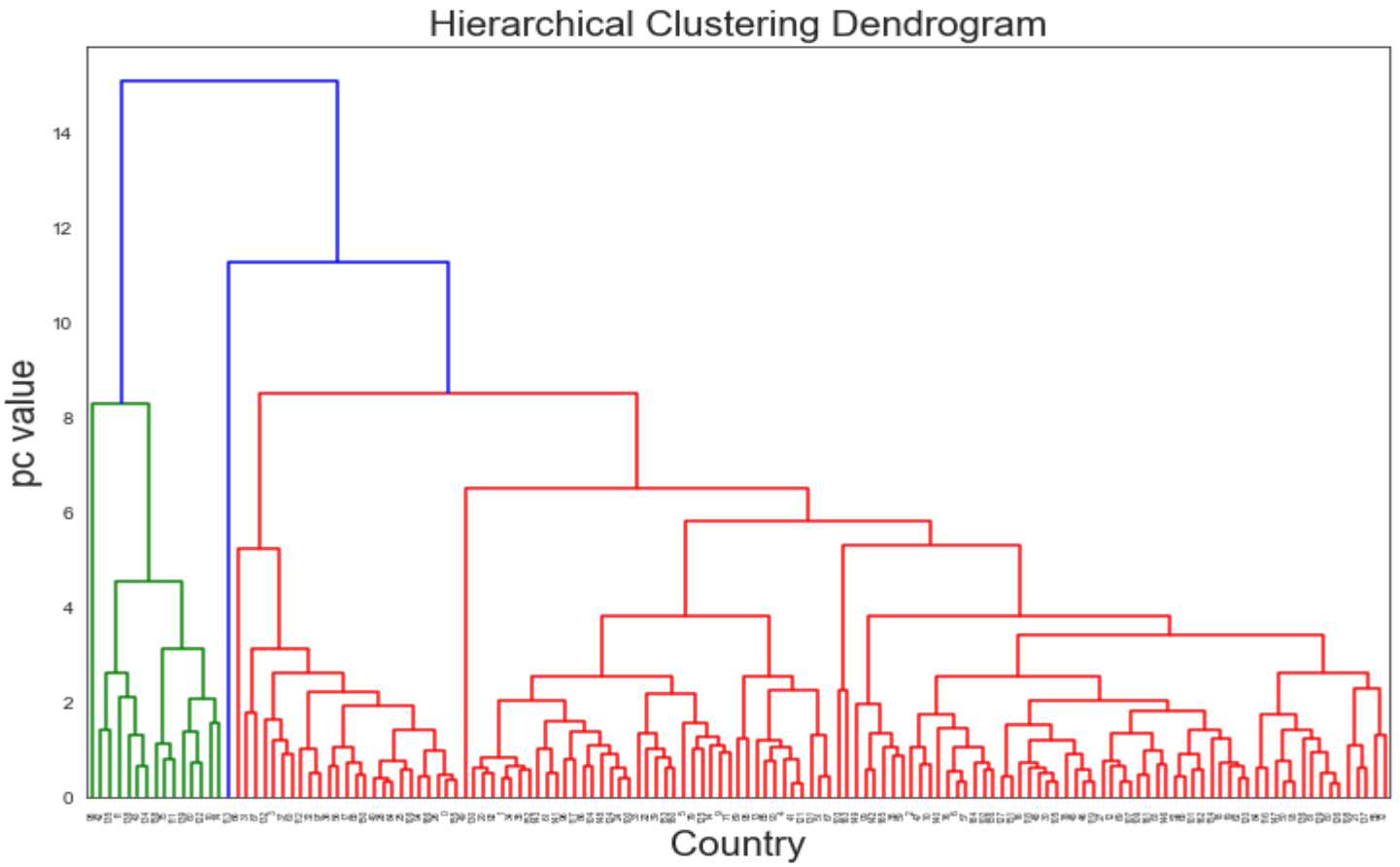
- It has highest child mortality
- Lowest income
- Highest Inflation
- Comparatively low life expectancy
- Highest total fertility
- Lowest gdpp



Hierarchical Clustering



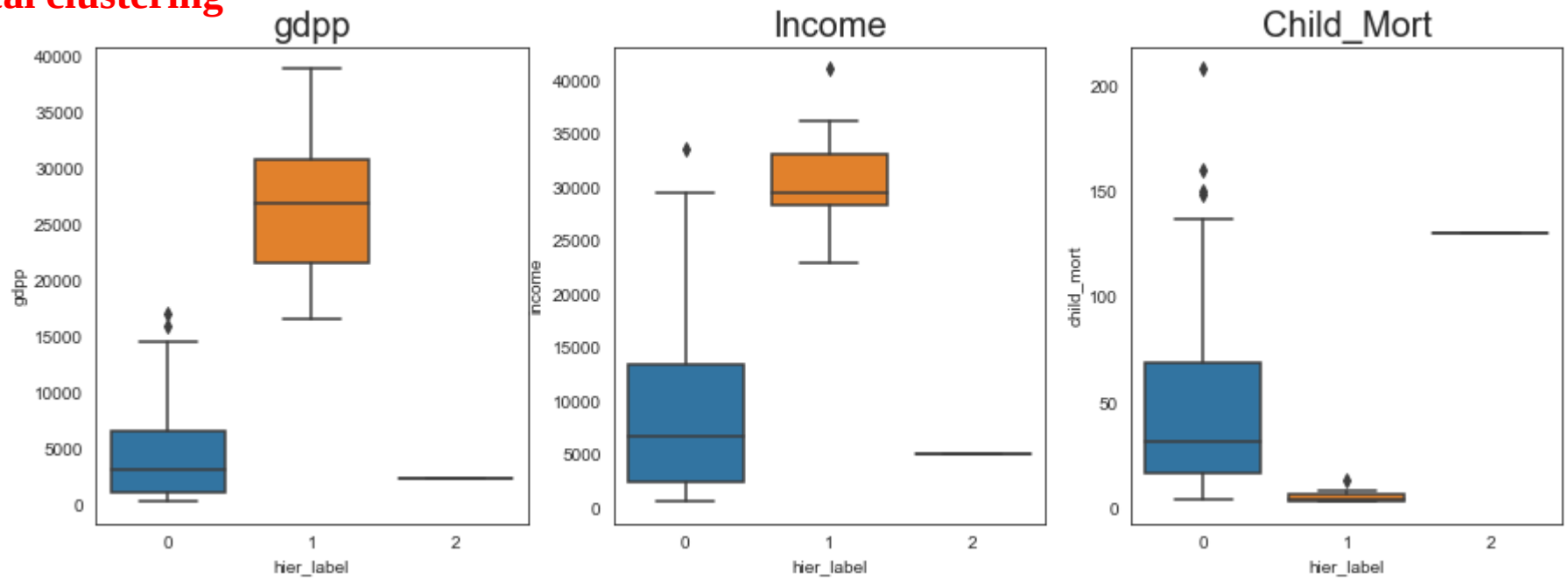
Single Linkage



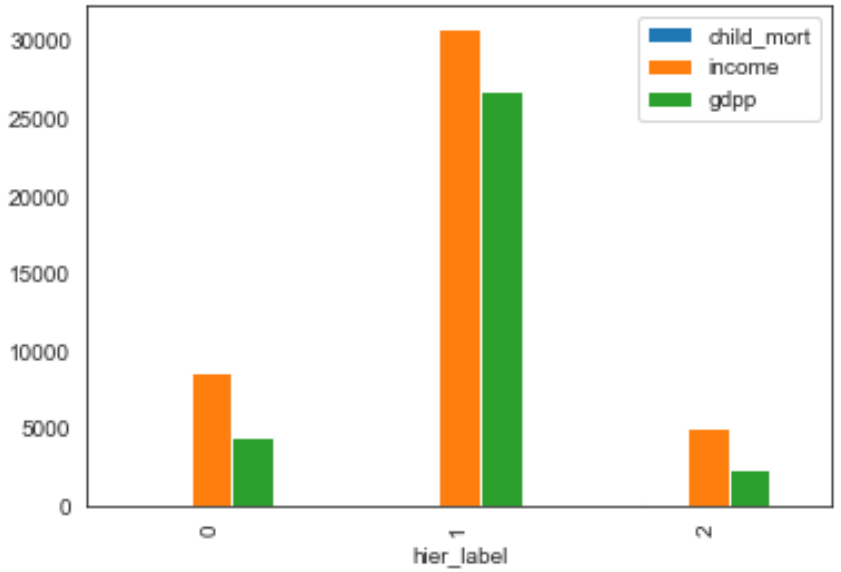
Complete Linkage

Taking 3 Clusters for analysis in Hierarchical clustering from complete linkage

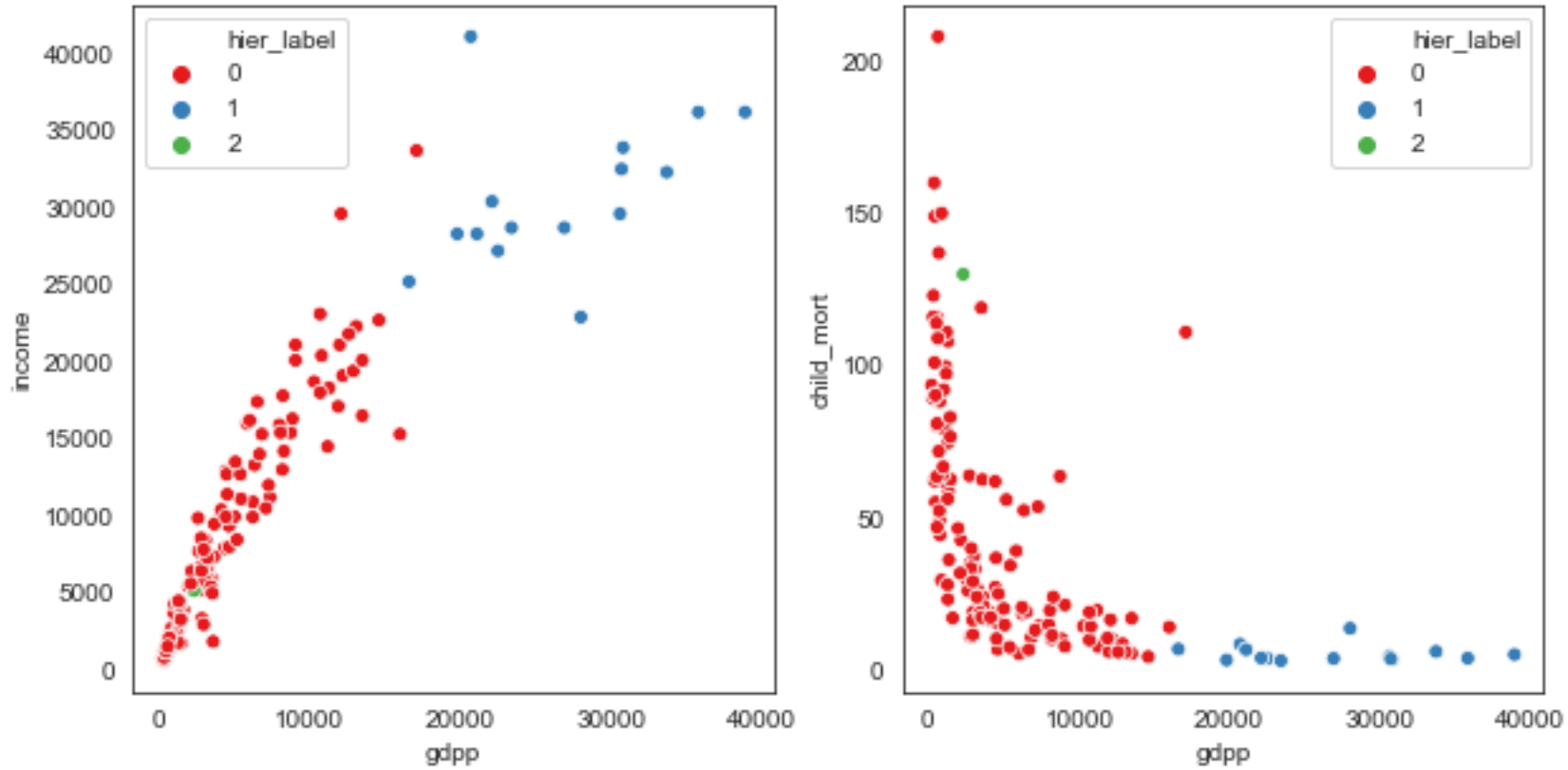
Cluster Profiling for Hierarchical clustering



- ✓ **For cluster 0:** gdpp and income is the Quite moderate , Mortality of children is highest.
- ✓ **For cluster 1:** gdpp and income is the highest among other clusters, Mortality of children is lowest.
- ✓ **For cluster 2:** gdpp and income is the lowest than other clusters, Mortality of children is very high than other clusters. but it is having only one countries

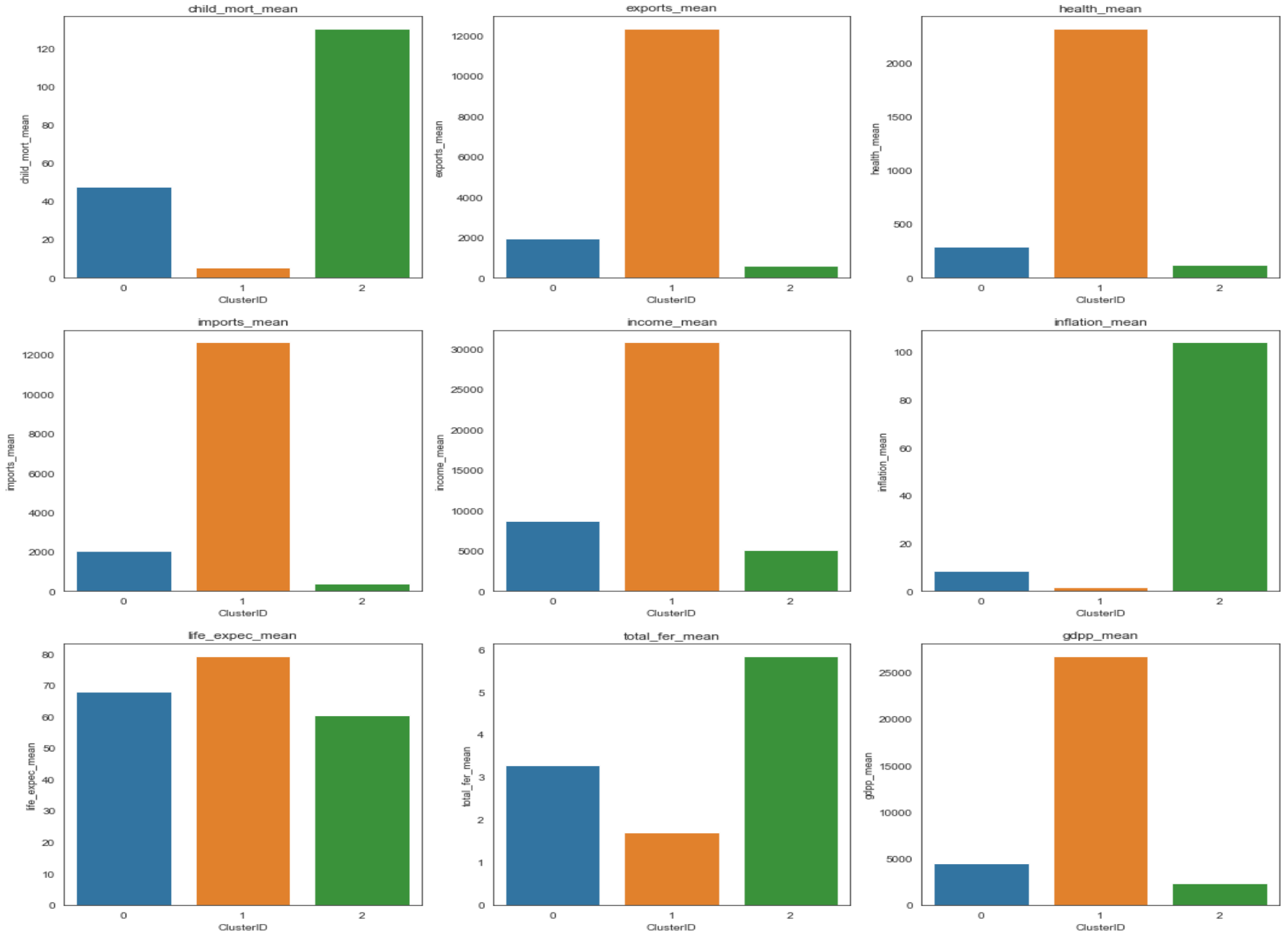


Clustering Pattern for Hierarchical clustering



Looking at the graph we are certain that cluster 0 is cluster of concern. Because it is having more countries with:

- It has highest child mortality
- Lowest income
- Highest Inflation
- Comparatively low life expectancy
- Highest total fertility
- Lowest gdpp



Final list of Top 10 Countries in Need of Aid

1. **Burundi**
2. **Liberia**
3. **Congo, Dem. Rep**
4. **Niger**
5. **Sierra Leone**
6. **Madagascar**
7. **Mozambique**
8. **Central African Republic**
9. **Malawi**
10. **Eritre**

CONCLUSION

In Hierarchical clustering it is showing 127 countries which are in need of aid while in K means clustering it is showing 46 countries. I would choose the final countries from hierarchical clustering as it gave accurate output than k-means clustering. I have compared the clusters and visualized from both methods and hierarchical clustering gave precise information than K-Means clustering.

Thank You