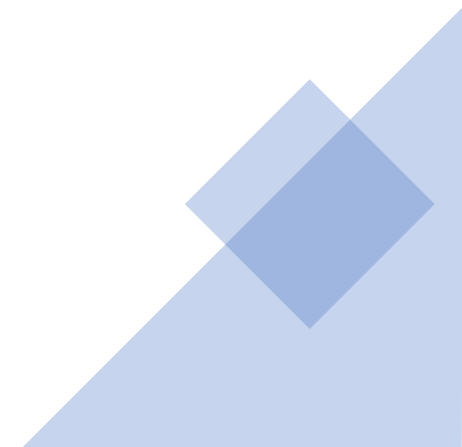




HELPIng determine countries  
in need of aid

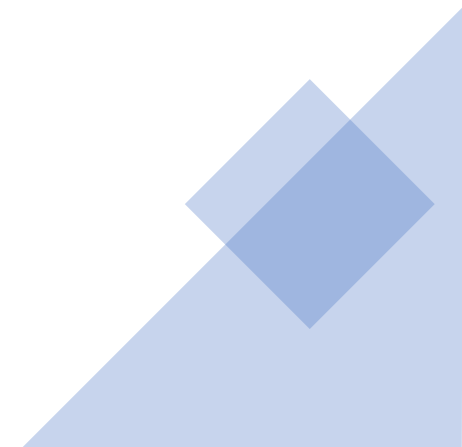


# Agenda

- Problem Statement
  - Understanding approach
  - Steps performed for analysis
  - Final recommendations
- 




# Problem Statement

- As part of distribution of funds raised, by HELP international, deciding on the countries most in need of aid.
  - Countries to be categorized as per socio-economic and health factors and overall development of countries to determine which countries need to be focussed upon.
- 


# Understanding the approach

- As part of the determination of countries in need, it was required to determine how the countries could be grouped. This was done using a modelling methodology called 'Clustering'.
- Clustering can be done in multiple ways. The methods applied here were the K-Means method and Hierarchical Clustering method.
- Both methods have their own advantages and disadvantages and the method best able to identify the clustering is chosen for the determination of final countries.



## Steps performed for analysis (1)

- Data Analysis:
    - As part of the analysis, determined that there was no missing data.
    - It was observed that some of the data was described as percentages rather than absolute values (exports, health and imports were described as percentage of GDPP) and were subsequently converted. This conversion gives more clarity about the actual situation since gdpp is vastly different across countries.

e.g. Venezuela has just 17.6% imports while Vietnam has 80.2% imports, however once we convert these to absolutes, Venezuela imports value is 2376 while that for Vietnam is just 1050.620
- 

# Steps performed for analysis (2)

- Data Correlation:
  - Checked for correlation amongst different variables using pair plots and heatmap

Looking at the pair plots, correlation matrix and heatmap, the following variables are highly correlated

imports and exports: 99%

gdpp and health: 92%

gdpp and income: 90%

total\_fer and child\_mort: 85%

gdpp and exports: 77%

gdpp and imports: 76%

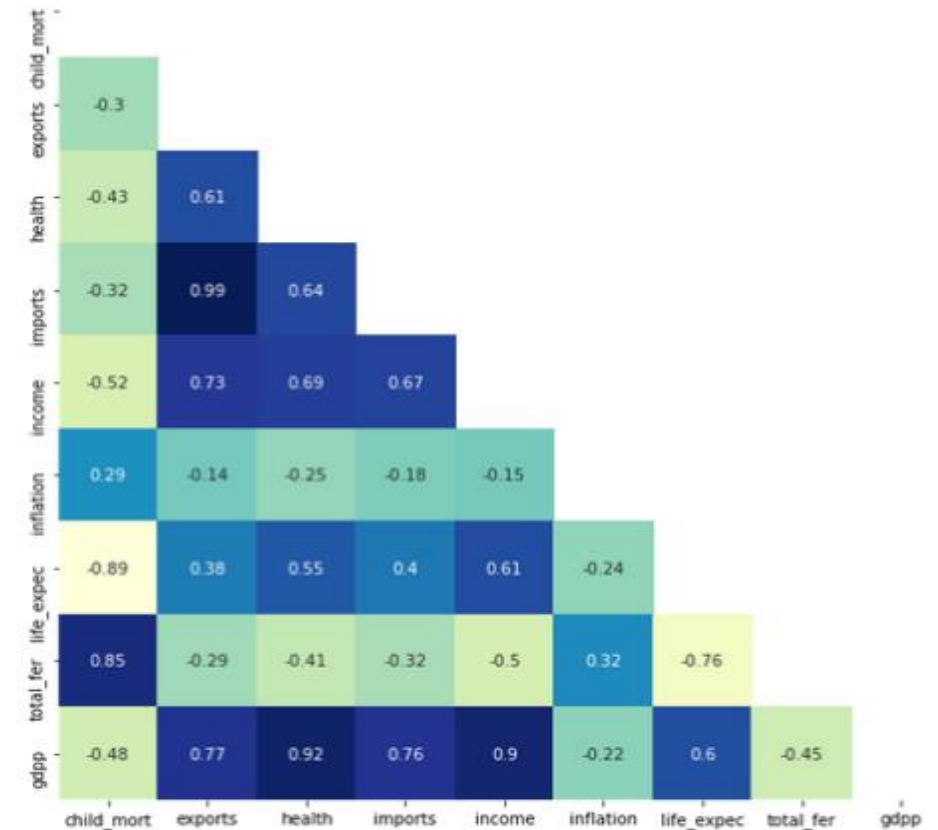
exports and income: 73%

In all the cases above, there is positive correlation between the variables i.e. as one variable increases so does the other.

life\_expec and child\_mort: 89%

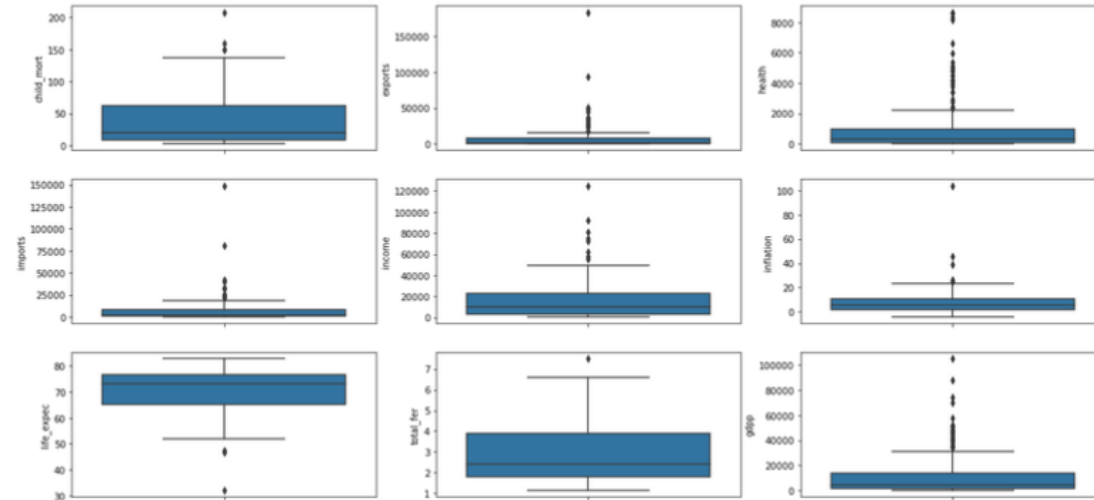
total\_fer and life\_expec: 76%

In the cases above, there is a negative correlation between the variables i.e. as one variable increases, the other decreases.



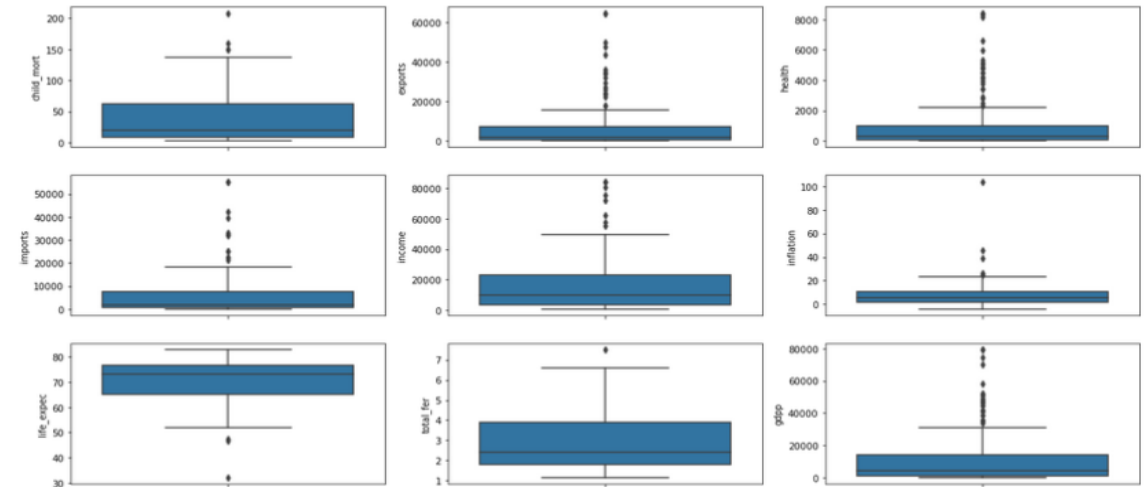
# Steps performed for analysis (3)

- Outlier Analysis and treatment:
  - It was observed that many of the variables had outliers.
  - Some of the variables like high child mortality, high inflation, lower life expectancy and high total fertility (i.e. number of children) typically impact underdeveloped countries, so no outliers were removed here.
  - Higher outliers for exports, imports, health investment, income and GDP typically relate more to developed countries and were capped to 99% values.



Before treatment

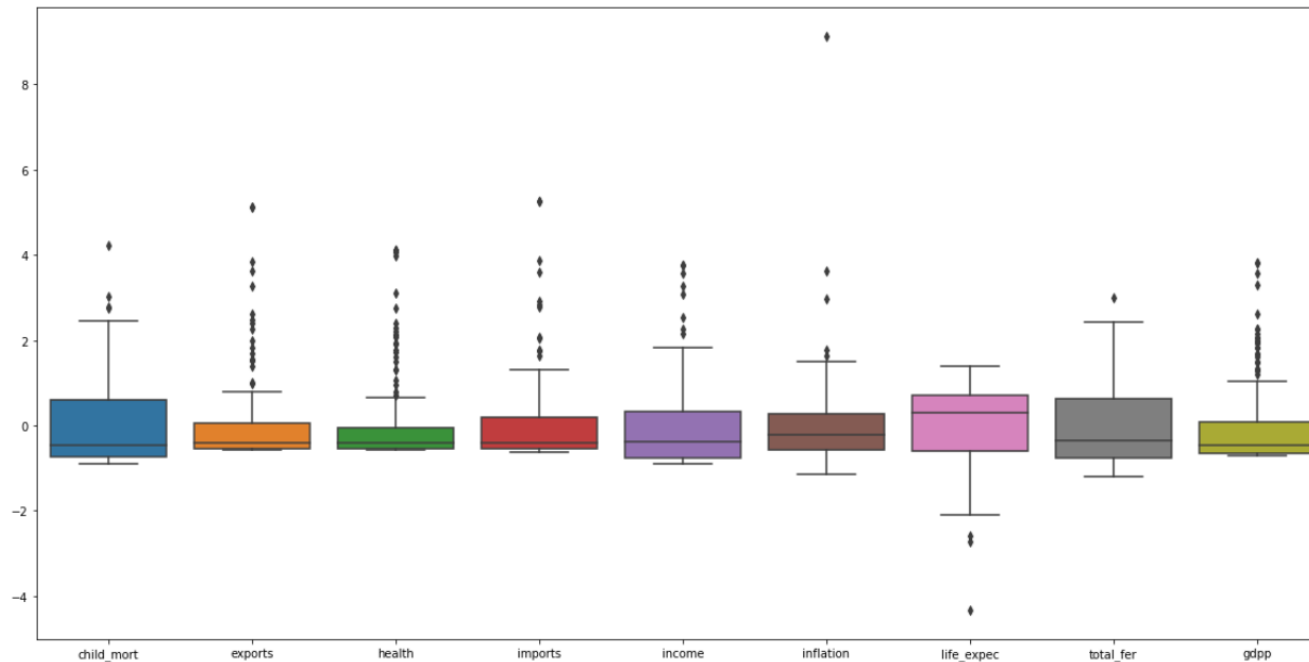
After treatment



# Steps performed for analysis (4)

Rescaling:

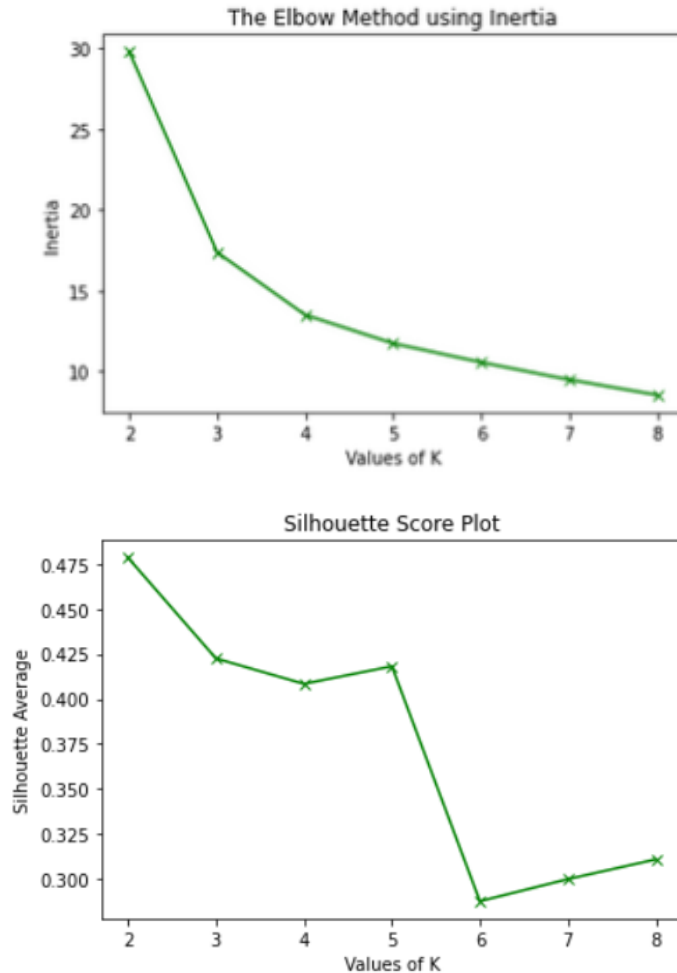
Data was scaled to a common scale for analysis.



- Determining cluster tendency:
  - Cluster tendency was determined using Hopkin's test. With a score of 0.92, it was determined to have a good cluster tendency.

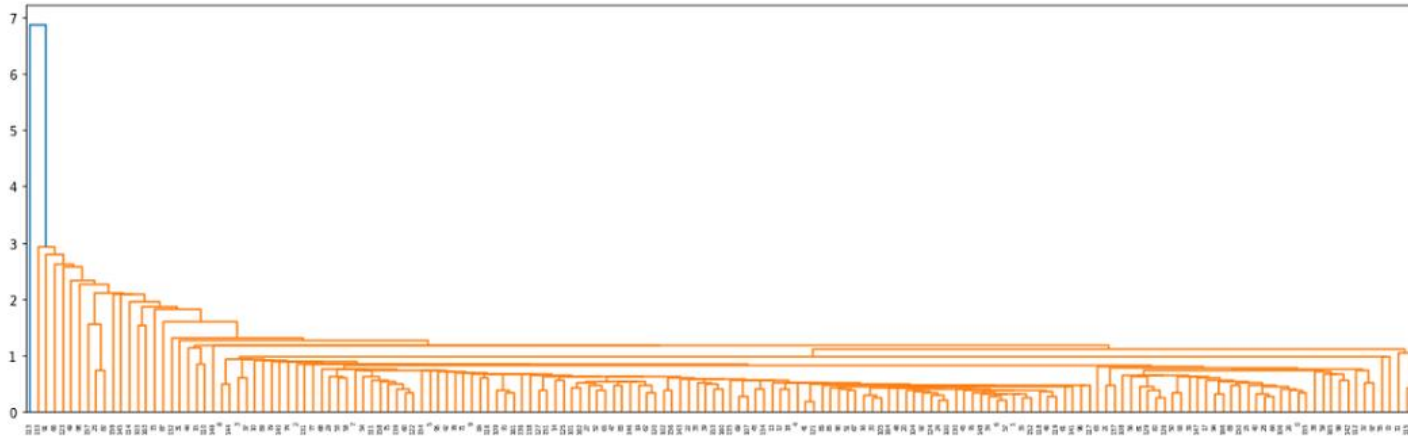


# Model Building (1)

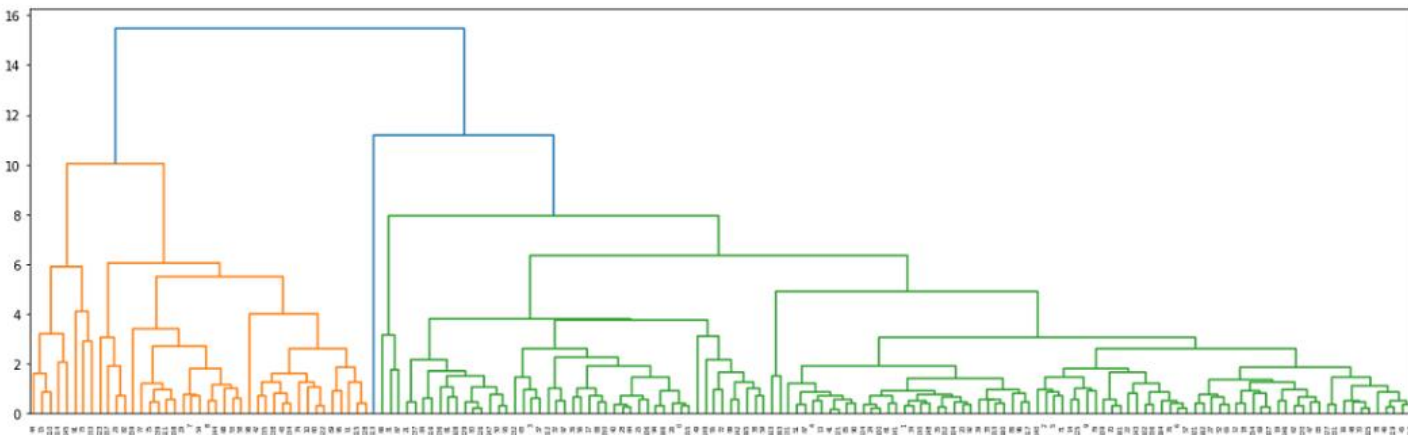


- 2 methods were employed to determine the best clustering possible to find the underdeveloped nations:
  - K-Means algorithm:
    - This method employs business defined cluster numbers to divide the data into different clusters.
    - Sum of Square Distances (SSD) or Elbow Curve Method and Silhouette Score were two methods used to decide the number of clusters. While SSD displayed 3 clusters as ideal, Silhouette score displayed a peak at 5 with a minor peak at 3. Both values were tested for K-Means and k=3 was found ideal.

# Model Building (2)



Single linkage clustering shows only 1 point in two of the clusters if we consider 3 clusters. Even increasing the cluster number displays the same kind of distribution. Hence, going for complete linkage method



Complete linkage method appears to display more distinct clusters. However, even here, one of the clusters appears to display on cluster member. We will check for different values of cluster numbers to look for distribution of cluster elements using complete linkage rather than single linkage

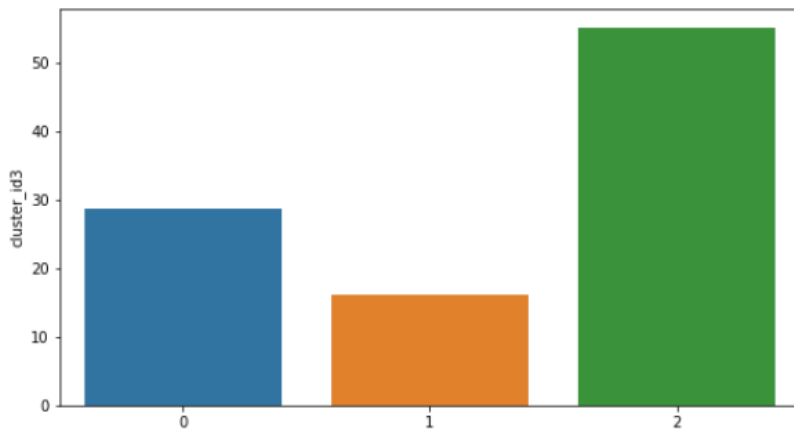
- Hierarchical clustering (Agglomerative technique):
  - This method keeps grouping nearby data points to come up with stable clusters which can then be split into multiple clusters as per business demands.
  - Two methods were used here to determine clusters – Single Linkage and Complete Linkage.

# Model Building (3)

- Following the two clustering techniques, the following cluster distributions were observed:

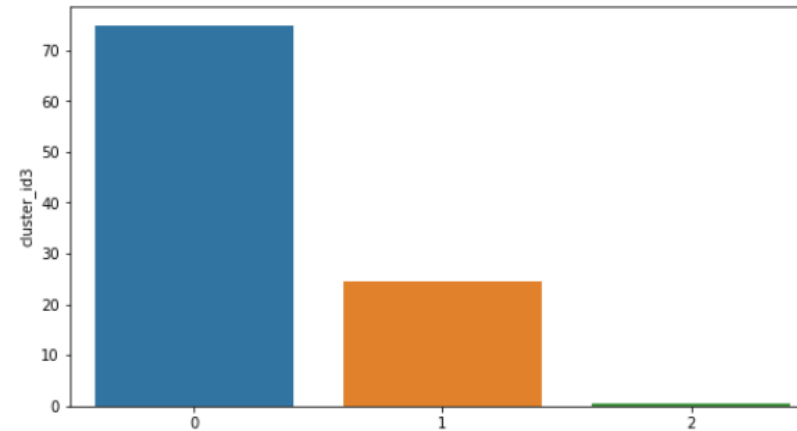
## K-Means Clustering:

- Cluster members here was distributed in a quite expected fashion when using 3 clusters with not a very high difference between largest and smallest clusters.



## Hierarchical clustering:

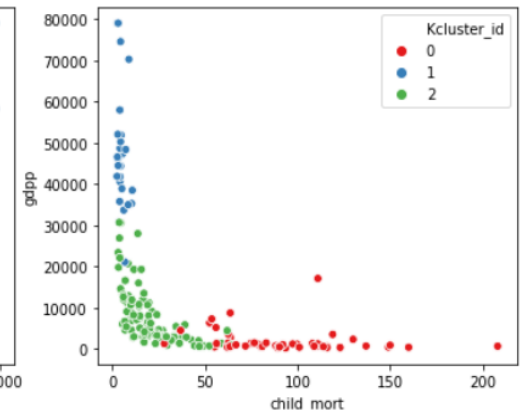
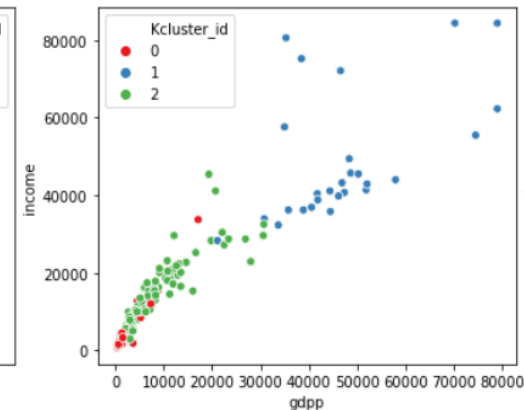
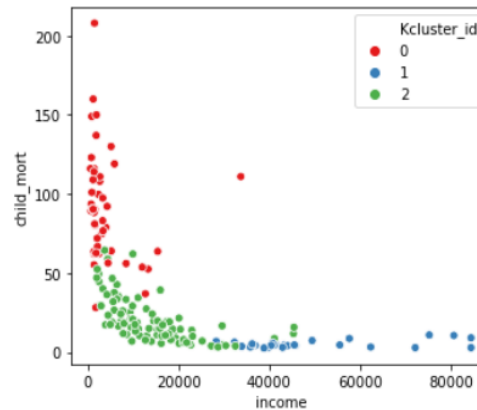
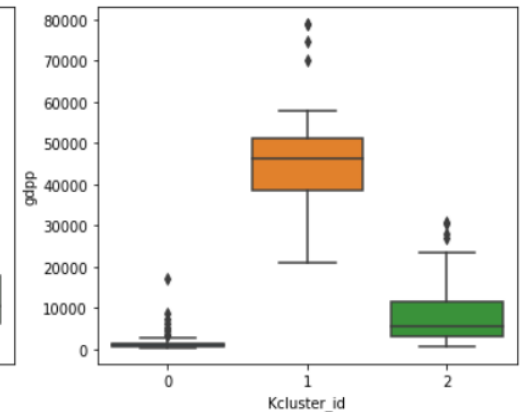
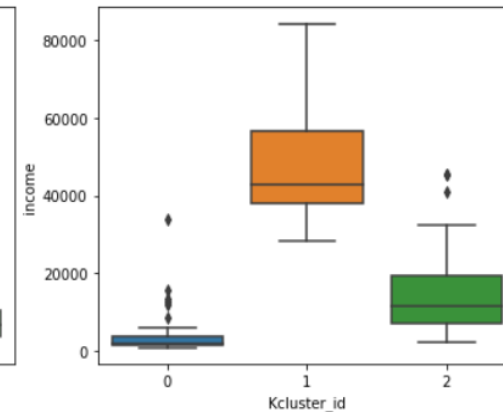
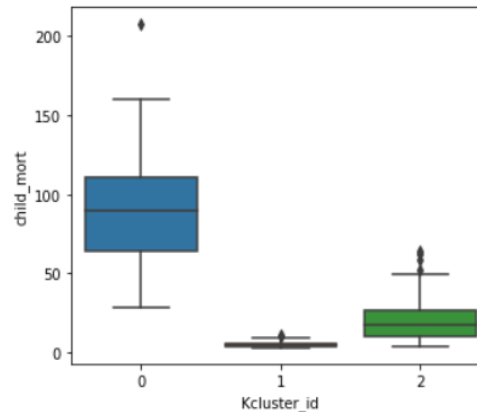
- The data was highly skewed towards the cluster with most data points with those with least data points being very less.



Accordingly, **K-Means clustering with 3 clusters** was determined as final modelling technique. This also helps divide clusters as Underdeveloped, Developing and Developed nation clusters.

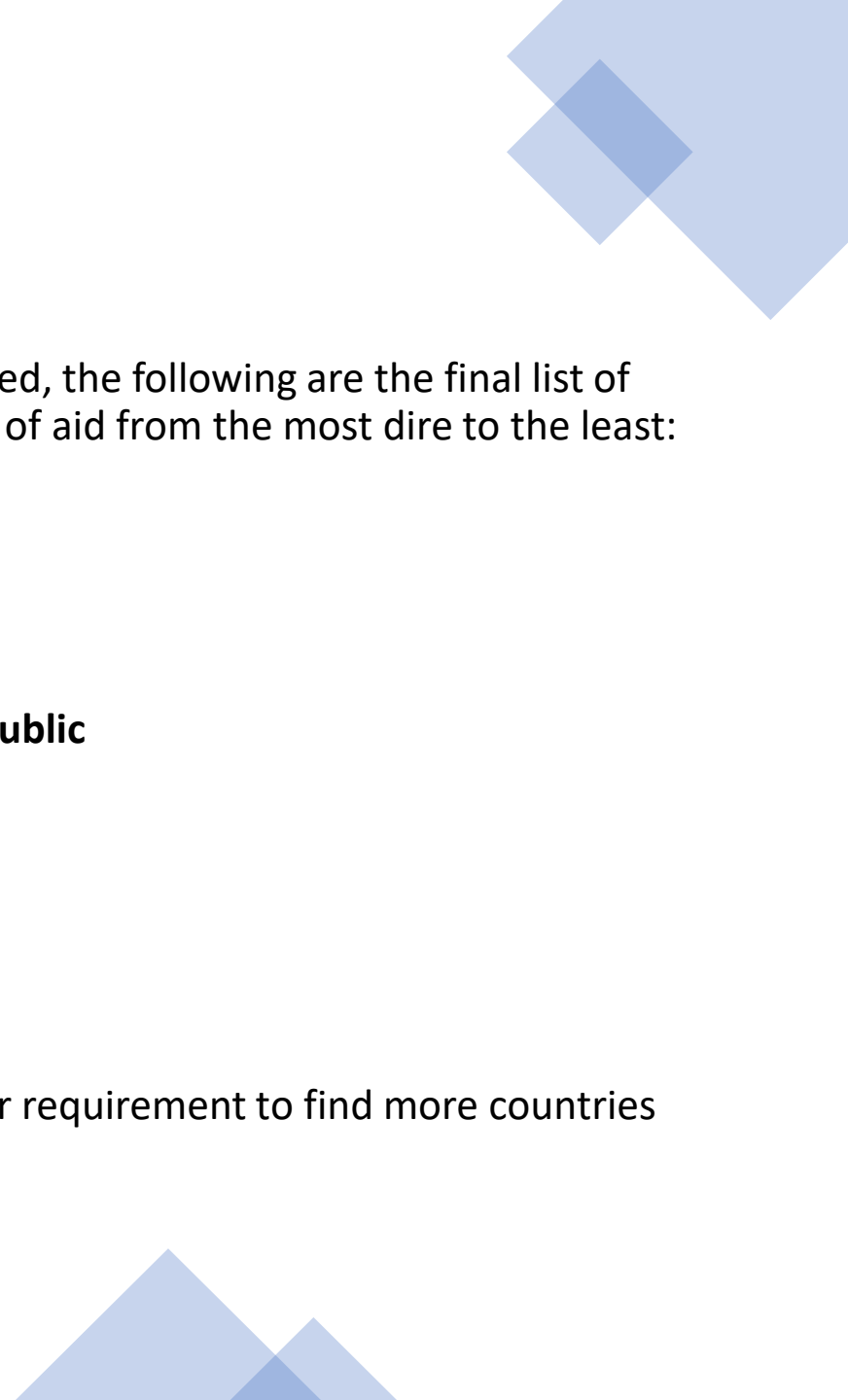
# Model Evaluation

- Plotting box plots helped determine that there is a reliable distinction between clusters observed.
- The scatter plots reveal that child mortality is higher where GDPP and Income are low and vice versa. The income and GDPP have a positive linear correlation.
- As determined earlier, underdeveloped nations are characterized by higher child mortality and lower income and GDPP. Hence, cluster 0 appeared most relevant.





# Final list of countries in need of aid

- As per the evaluation performed, the following are the final list of top 10 countries most in need of aid from the most dire to the least:
    - 1. Congo, Dem. Rep.**
    - 2. Liberia**
    - 3. Burundi**
    - 4. Niger**
    - 5. Central African Republic**
    - 6. Mozambique**
    - 7. Malawi**
    - 8. Guinea**
    - 9. Togo**
    - 10. Sierra Leone**
  - The list can be extended as per requirement to find more countries in need.
- 



Thank You