
# CLUSTERING ASSIGNMENT

BY- MANALI GHOSH


# PROBLEM STATEMENT

## BACKGROUND

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.

## CONTEXT

After the recent funding programs, 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.

## PROBLEM

The job is to categories the countries using some socio-economic and health factors that determine the overall development of the country. Then the countries which the CEO needs to focus on the most needs to be analyzed and suggested.

# STEP WISE APPROACH

1. Data Quality Check
2. EDA: Univariate and Bivariate Analysis
3. Outlier
4. Hopkin's Test
5. Scaling
6. Finding best value ok k: Silhouette Score and SSD elbow
7. Final K mean Analysis
8. Visualization of the clusters using Scatterplot
9. Cluster profiling: GDP, child mortality and income
10. Hierarchial analysis: Single linkage and Complete linkage and visualization
11. Final list of countries with insights


# DATA QUALITY CHECK

| Column Name | Description |
|---|---|
| country | Name of the country |
| child_mort | Death of children under 5 years of age per 1000 live births |
| exports | Exports of goods and services per capita. Given as %age of the GDP per capita |
| health | Total health spending per capita. Given as %age of GDP per capita |
| imports | Imports of goods and services per capita. Given as %age of the GDP per capita |
| Income | Net income per person |
| Inflation | The measurement of the annual growth rate of the Total GDP |
| life_expec | The average number of years a new born child would live if the current mortality patterns are to remain the same |
| total_fer | The number of children that would be born to each woman if the current age-fertility rates remain the same. |
| gdpp | The GDP per capita. Calculated as the Total GDP divided by the total population. |

| | country | child_mort | exports | health | imports | income | inflation | life_expec | total_fer | gdpp |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Afghanistan | 90.2 | 10.0 | 7.58 | 44.9 | 1610 | 9.44 | 56.2 | 5.82 | 553 |
| 1 | Albania | 16.6 | 28.0 | 6.55 | 48.6 | 9930 | 4.49 | 76.3 | 1.65 | 4090 |
| 2 | Algeria | 27.3 | 38.4 | 4.17 | 31.4 | 12900 | 16.10 | 76.5 | 2.89 | 4460 |
| 3 | Angola | 119.0 | 62.3 | 2.85 | 42.9 | 5900 | 22.40 | 60.1 | 6.16 | 3530 |
| 4 | Antigua and Barbuda | 10.3 | 45.5 | 6.03 | 58.9 | 19100 | 1.44 | 76.8 | 2.13 | 12200 |

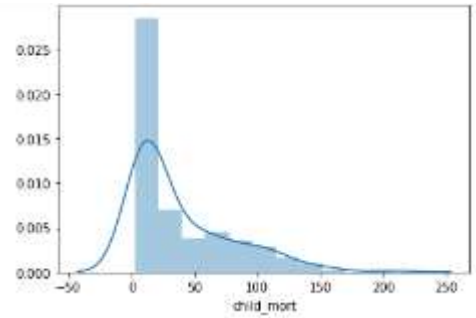
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 167 entries, 0 to 166
Data columns (total 10 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   country     167 non-null    object
 1   child_mort  167 non-null    float64
 2   exports     167 non-null    float64
 3   health      167 non-null    float64
 4   imports     167 non-null    float64
 5   income      167 non-null    int64
 6   inflation   167 non-null    float64
 7   life_expec  167 non-null    float64
 8   total_fer   167 non-null    float64
 9   gdpp        167 non-null    int64
dtypes: float64(7), int64(2), object(1)
memory usage: 13.2+ KB
```

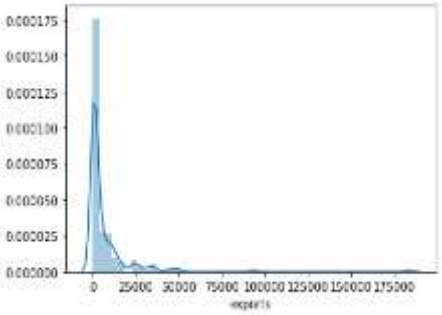
INSIGHTS:

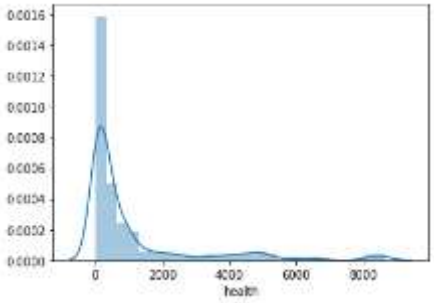
1. The dataset consists of 167 rows and 10 columns
2. There are no null or missing values

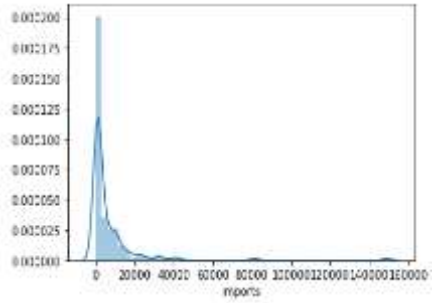
# EDA- UNIVARIATE ANALYSIS

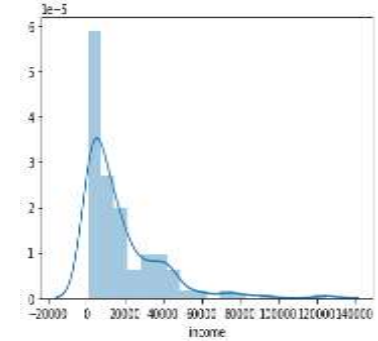
A dist. plot is made for the below columns in the dataset

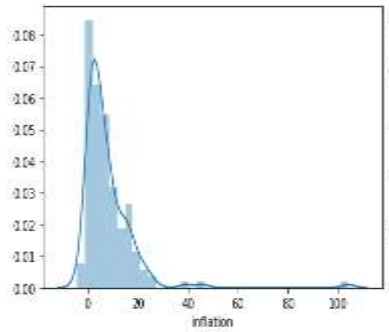


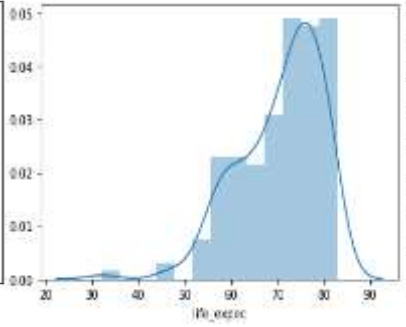



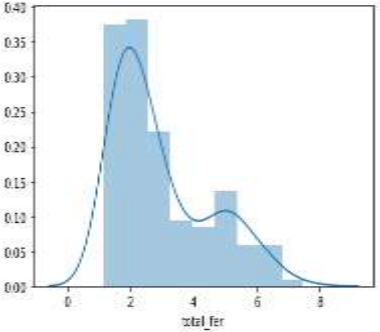



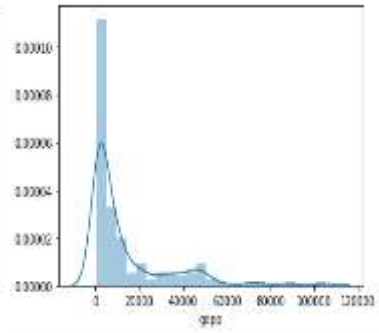



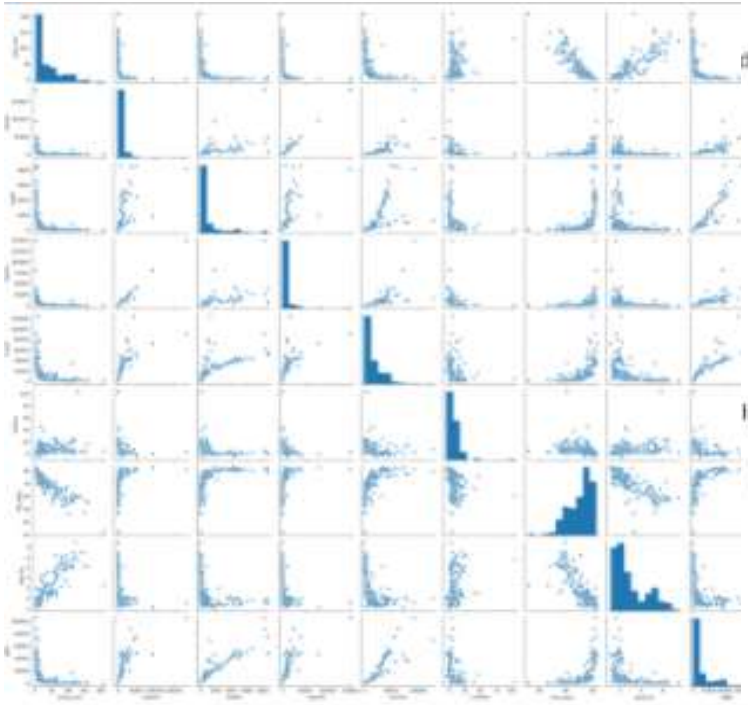



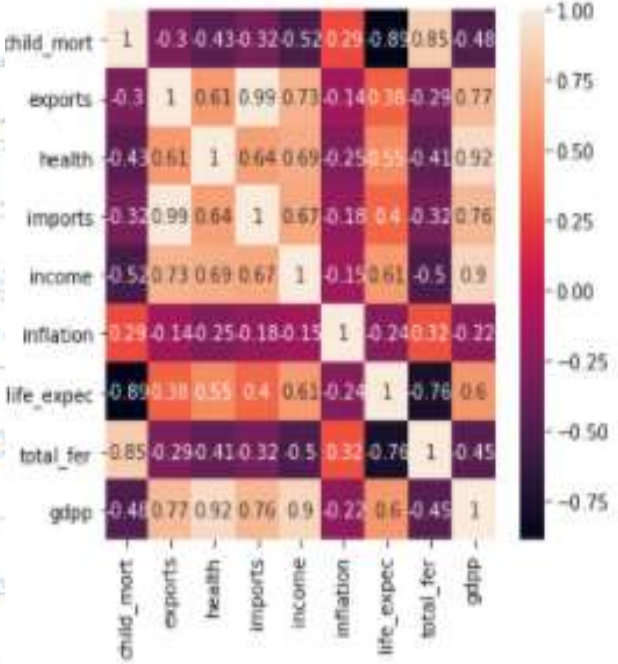



# EDA- BIVARIATE ANALYSIS

A heat map and a pair plot is plotted between the parameters in order to understand the correlation






INSIGHTS:

1. We can see high negative correlation between life expectancy and child mortality, and in between life expectancy and total fertility

2. We can see high correlation between exports and imports, and child mortality and total fertility

3. exports, imports, health and income displays a good positive correlation with GDP per capita

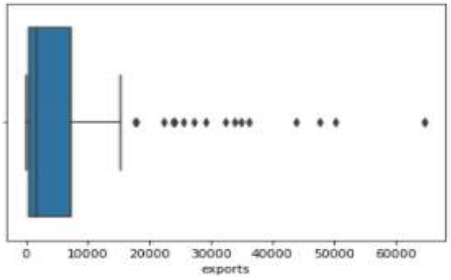
# OUTLIER TREATMENT

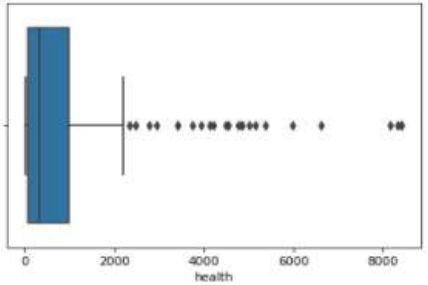
- We see a high amount of outliers in the data, and we need to implement the necessary outlier treatment
- We use soft capping outlier treatment in this case, as the data size is quite small. So eliminating rows can lead to a erroneous model
- SOFT CAPPING :

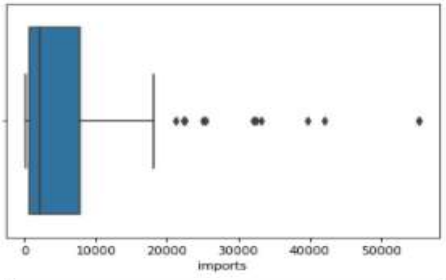
  a. If data between 1-99 % Keep as it is

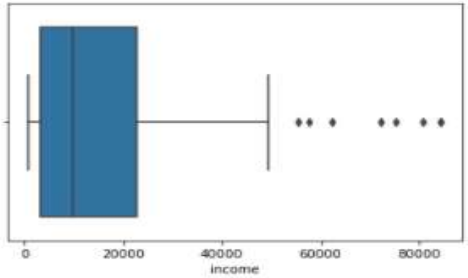
  b. value <1%- Bring data to 1%

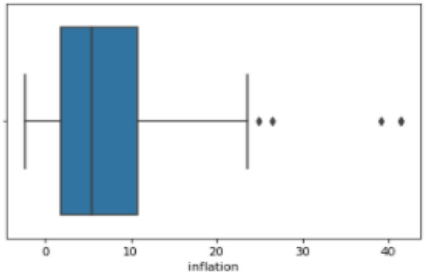
  c. value >99%- Bring data to 99%

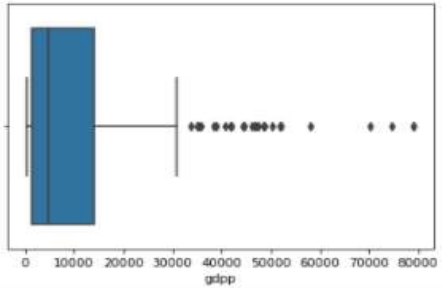



INSIGHTS:

1. We use soft capping in order to prevent dropping on any rows which can impact our analysis

2. We have not capped any lower end outliers as these variables could be from those countries which require aid from the NGO

3. We have mostly capped outliers to 99 percentile as those countries are categorized as developed nations and will not require external aids from the NGO


# HOPKIN'S TEST

The Hopkins statistic, is a statistic which gives a value which indicates the cluster tendency, in other words: how well the data can be clustered.

If the value is between {0.01, ...,0.3}, the data is regularly spaced.
If the value is around 0.5, it is random.
If the value is between {0.65, ..., 0.84}, it is good data for clustering
If the value is between {0.85 and above}, it has a high tendency to cluster.

Here our Hopkin's test score is 0.92, so we can definitely use this data for clustering

```
1 hopkins(country_df.drop('country', axis=1))
```

```
0.9292563888865134
```


# SCALING

**Scaling** or Standardization: It is a step of Data Pre Processing which is applied to independent variables or features of data. It basically helps to normalize the data within a particular range. Sometimes, it also helps in speeding up the calculations in an algorithm.

Here we are using the standard scalar.

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 1.344012 | -0.569638 | -0.566983 | -0.598844 | -0.851772 | 0.263649 | -1.619092 | 1.902882 | -0.702314 |
| 1 | -0.547543 | -0.473873 | -0.440417 | -0.413679 | -0.387025 | -0.375251 | 0.647866 | -0.859973 | -0.498775 |
| 2 | -0.272548 | -0.424015 | -0.486295 | -0.476198 | -0.221124 | 1.123260 | 0.670423 | -0.038404 | -0.477483 |
| 3 | 2.084186 | -0.381264 | -0.534113 | -0.464070 | -0.612136 | 1.936405 | -1.179234 | 2.128151 | -0.531000 |
| 4 | -0.709457 | -0.086754 | -0.178431 | 0.139659 | 0.125202 | -0.768917 | 0.704258 | -0.541946 | -0.032079 |


# FINDING THE BEST VALUE OF K- Silhouette SCORE

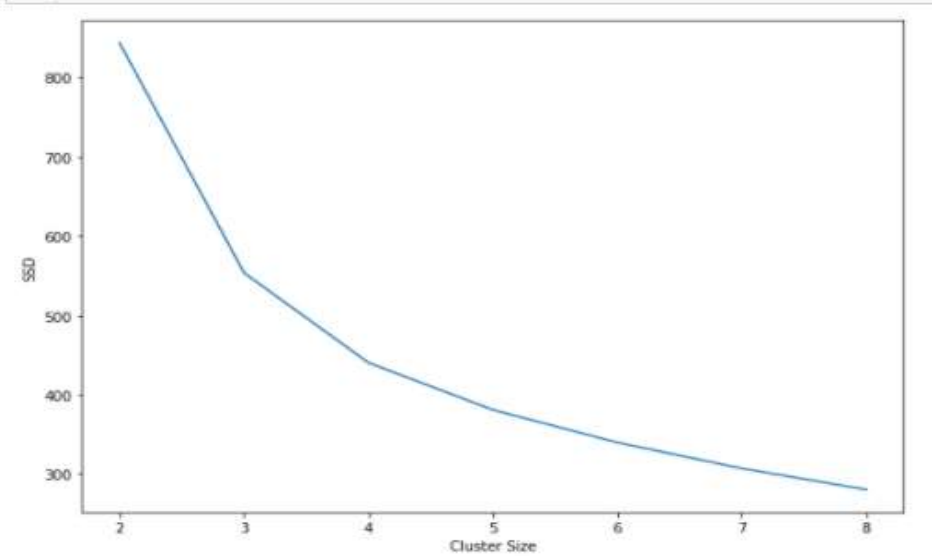
Silhouette Analysis silhouette score=$p-qmax(p,q)$
$p$ is the mean distance to the points in the nearest cluster that the data point is not a part of
$q$ is the mean intra-cluster distance to all the points in its own cluster.
The value of the silhouette score range lies between -1 to 1.
A score closer to 1 indicates that the data point is very similar to other data points in the cluster,
A score closer to -1 indicates that the data point is not similar to the data points in its cluster.

```
For n_clusters=2, the silhouette score is 0.4691402080700722
For n_clusters=3, the silhouette score is 0.40442323247838274
For n_clusters=4, the silhouette score is 0.39278006155982664
For n_clusters=5, the silhouette score is 0.38414464865853404
For n_clusters=6, the silhouette score is 0.2957589757167241
For n_clusters=7, the silhouette score is 0.30770914718834236
For n_clusters=8, the silhouette score is 0.2773350207329115
```

INSIGHTS:
The sihouette score drops after K = 3 Based on above analysis, 3 seems to be the optimum number to create clusters

# FINDING THE BEST VALUE OF K- ELBOW CURVE

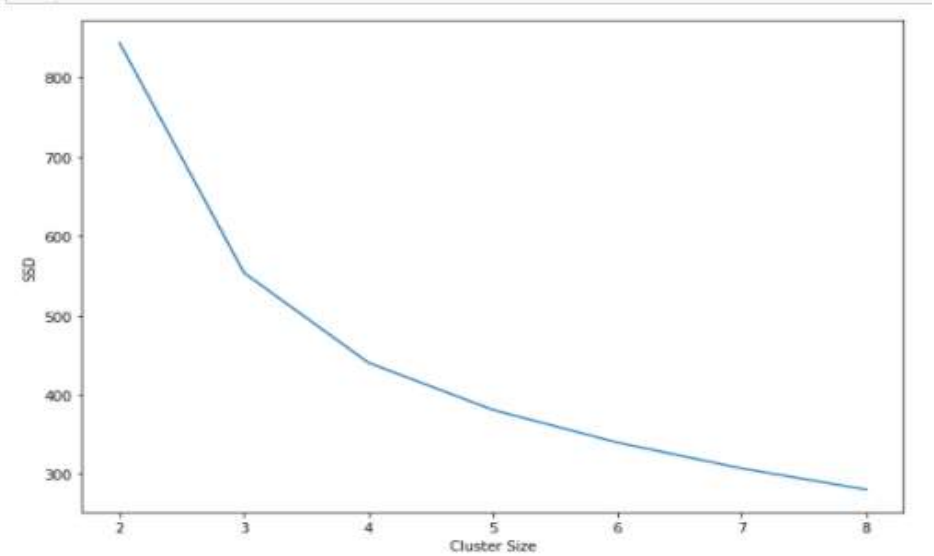
K-means clustering is an unsupervised learning algorithm which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest centroid. The algorithm aims to minimize the squared Euclidean distances between the observation and the centroid of cluster to which it belongs.
The elbow method helps to choose the optimum value of ‘k’ (number of clusters) by fitting the model with a range of values of ‘k’.




INSIGHTS:
Elbow curve - Post cluster N=3, there is not a very significant difference in distance between the points

# FINDING THE BEST VALUE OF K- ELBOW CURVE

K-means clustering is an unsupervised learning algorithm which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest centroid. The algorithm aims to minimize the squared Euclidean distances between the observation and the centroid of cluster to which it belongs.
The elbow method helps to choose the optimum value of ‘k’ (number of clusters) by fitting the model with a range of values of ‘k’.




INSIGHTS:
Elbow curve - Post cluster N=3, there is not a very significant difference in distance between the points


# FINAL K MEAN ANALYSIS

- **K Means** algorithm is an iterative algorithm that tries to partition the dataset into *K*pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to **only one group**.
- It tries to make the intra-cluster data points as similar as possible while also keeping the clusters as different (far) as possible.
- It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum.

```
3 print(kmeans.cluster_centers_)
```

```
[[ 1.34749254 -0.49725446 -0.52608487 -0.53720606 -0.72379806  0.41448121
  -1.27627565  1.3543418  -0.6241439 ]
 [-0.44608365 -0.26896584 -0.29423633 -0.24319265 -0.17060836 -0.03429503
   0.32004093 -0.46744568 -0.28478671]
 [-0.84593495  1.65776344  1.78390839  1.64390447  1.72748481 -0.57960501
   1.11922578 -0.79097569  1.91688659]]
```

```
1 kmeans.labels_
```

```
array([0, 1, 1, 0, 1, 1, 1, 2, 2, 1, 1, 1, 1, 1, 1, 2, 1, 0, 1, 1, 1, 0,
       1, 2, 1, 0, 0, 1, 0, 2, 1, 0, 0, 1, 1, 1, 0, 0, 0, 1, 0, 1, 2, 1,
       2, 1, 1, 1, 1, 0, 0, 1, 1, 2, 2, 0, 0, 1, 2, 0, 1, 1, 1, 0, 0, 1,
       0, 1, 2, 1, 1, 1, 0, 2, 1, 2, 1, 2, 1, 1, 0, 0, 2, 1, 0, 1, 1, 0,
       0, 1, 1, 2, 1, 0, 0, 1, 1, 0, 2, 0, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1,
       2, 2, 0, 0, 2, 1, 0, 1, 1, 1, 1, 1, 1, 2, 1, 1, 0, 1, 1, 0, 1, 1,
       0, 2, 1, 2, 0, 0, 1, 2, 1, 1, 0, 1, 2, 2, 1, 0, 1, 0, 0, 1, 1, 1,
       1, 0, 1, 2, 2, 2, 1, 1, 1, 1, 1, 0, 0])
```

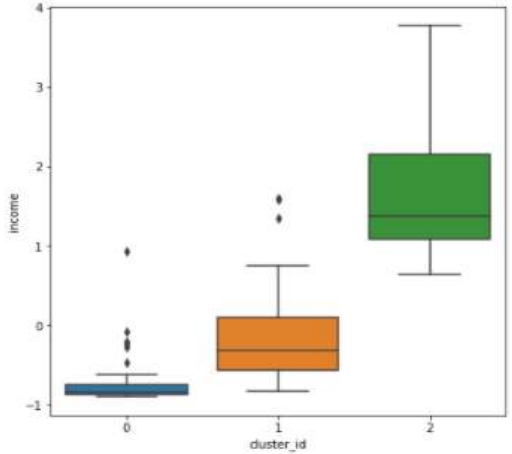
| | country | child_mort | exports | health | imports | income | inflation | life_expec | total_fer | gdpp | cluster_id |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Afghanistan | 1.344012 | -0.569638 | -0.566983 | -0.598844 | -0.851772 | 0.263649 | -1.619092 | 1.902882 | -0.702314 | 0 |
| 1 | Albania | -0.547543 | -0.473873 | -0.440417 | -0.413679 | -0.387025 | -0.375251 | 0.647866 | -0.859973 | -0.498775 | 1 |
| 2 | Algeria | -0.272548 | -0.424015 | -0.486295 | -0.476198 | -0.221124 | 1.123260 | 0.670423 | -0.038404 | -0.477483 | 1 |
| 3 | Angola | 2.084186 | -0.381264 | -0.534113 | -0.464070 | -0.612136 | 1.936405 | -1.179234 | 2.128151 | -0.531000 | 0 |
| 4 | Antigua and Barbuda | -0.709457 | -0.086754 | -0.178431 | 0.139659 | 0.125202 | -0.768917 | 0.704258 | -0.541946 | -0.032079 | 1 |

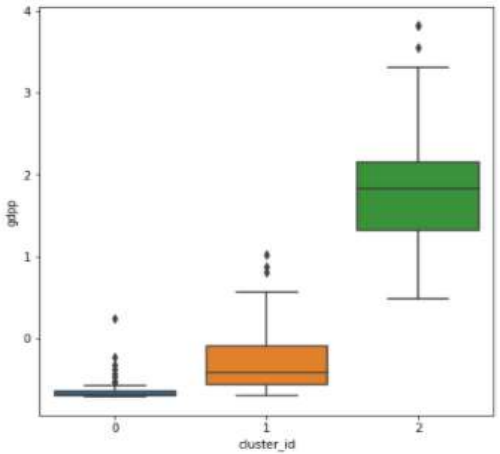
## Cluster details

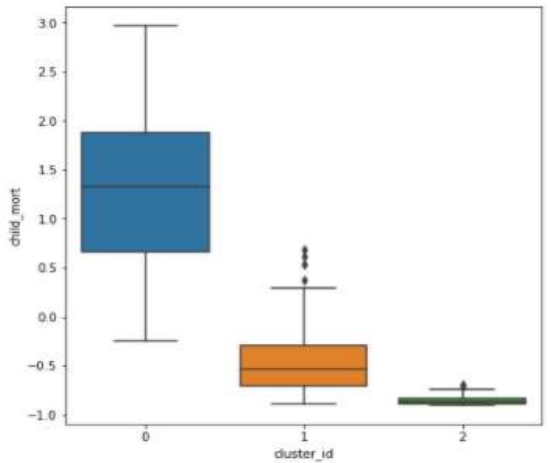
```
1    90
0    48
2    29
Name: cluster_id, dtype: int64
```

# VISUALIZATION OF CLUSTERS USING SCATTERPLOT

- Plotting a box plot with GDP per person, income and mortality in the y axis and Cluster_id in the x axis




INSIGHTS:

Cluster 0: Countries with low income have low GDP per person and high child mortality rate

Cluster 1: Countries with average income have average GDP per person and average mortality rate

Cluster 2: Countries with high income have high GDP per person and low child mortality rate

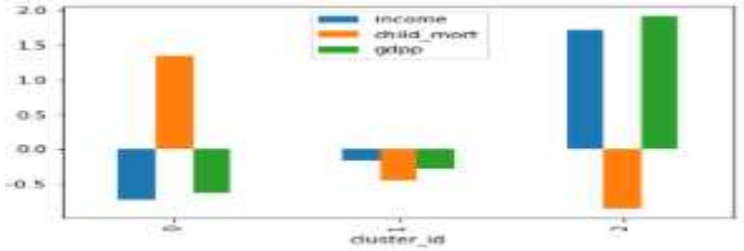
Therefore we should look into Cluster 0 countries for the NGO to provide aid to them

# CLUSTER PROFILING

The table below gives us the mean of GDP per person, income and child mortality per cluster wise

| cluster_id | income | child_mort | gdpp |
|---|---|---|---|
| 0 | -0.723798 | 1.347493 | -0.624144 |
| 1 | -0.170608 | -0.446084 | -0.284787 |
| 2 | 1.727485 | -0.845935 | 1.916887 |

The graph below shows the relationship between Cluster0, Cluster1 and Cluster2 in terms of income, GDP and child mortality



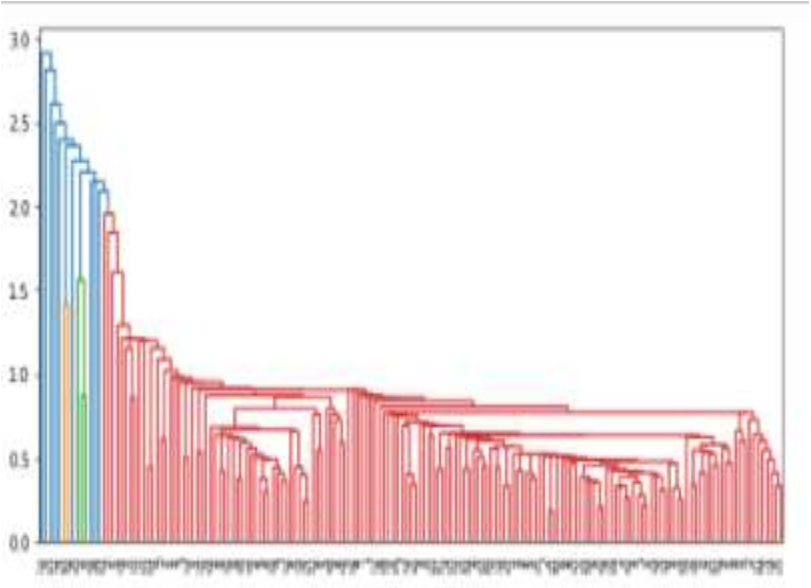

| | country | child_mort | exports | health | imports | income | inflation | life_expec | total_fer | gdpp | cluster_id |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 88 | Liberia | 1.320882 | -0.569009 | -0.568849 | -0.593042 | -0.900244 | -0.248762 | -1.100286 | 1.372838 | -0.715054 | 0 |
| 26 | Burundi | 1.431394 | -0.572543 | -0.575452 | -0.614108 | -0.899028 | 0.632791 | -1.449916 | 2.194407 | -0.715054 | 0 |
| 37 | Congo, Dem. Rep. | 2.007085 | -0.562435 | -0.575663 | -0.607641 | -0.900244 | 1.729892 | -1.472473 | 2.379922 | -0.714917 | 0 |
| 112 | Niger | 2.186988 | -0.567709 | -0.580403 | -0.607087 | -0.896235 | -0.625649 | -1.325854 | 3.009349 | -0.714111 | 0 |
| 132 | Sierra Leone | 2.968283 | -0.568607 | -0.561185 | -0.610622 | -0.873557 | 1.265237 | -1.754433 | 1.492098 | -0.711176 | 0 |
| 93 | Madagascar | 0.624399 | -0.565425 | -0.580933 | -0.606371 | -0.864061 | 0.179753 | -1.100286 | 1.094565 | -0.710370 | 0 |
| 106 | Mozambique | 1.621578 | -0.562900 | -0.578234 | -0.604669 | -0.890426 | 0.031321 | -1.810825 | 1.730618 | -0.710025 | 0 |
| 31 | Central African Republic | 2.855201 | -0.569873 | -0.580518 | -0.612694 | -0.892102 | -0.695347 | -2.600313 | 1.498724 | -0.708471 | 0 |
| 94 | Malawi | 1.351722 | -0.565302 | -0.573519 | -0.608223 | -0.884170 | 0.606977 | -1.968722 | 1.564979 | -0.707723 | 0 |
| 50 | Eritrea | 0.444495 | -0.572468 | -0.580933 | -0.613321 | -0.862385 | 0.542441 | -0.998780 | 1.101191 | -0.706400 | 0 |

INSIGHTS:

1. We can clearly see Cluster 0 has low income, low GDP and high child mortality, hence external aid should be provided to them
2. Here we see a list of 10 nations where the income and GDP are the lowest and child mortality is the highest
3. These nations needs grants from the NGO

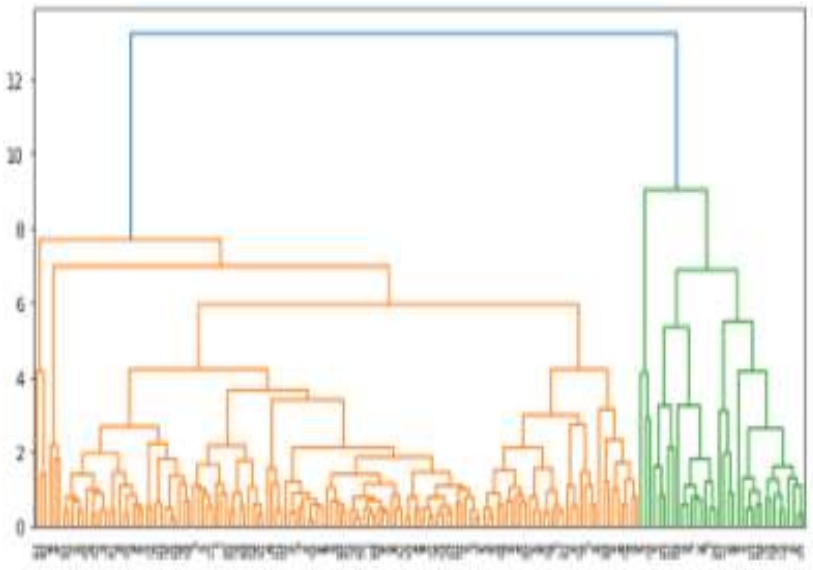
# HIERARCHIAL CLUSTERING

## Single linkage




INSIGHT:
The clusters obtained cannot lead to any conclusive outcomes

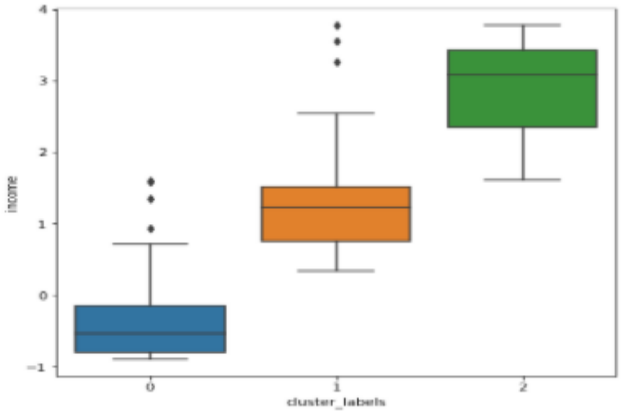
## Complete linkage

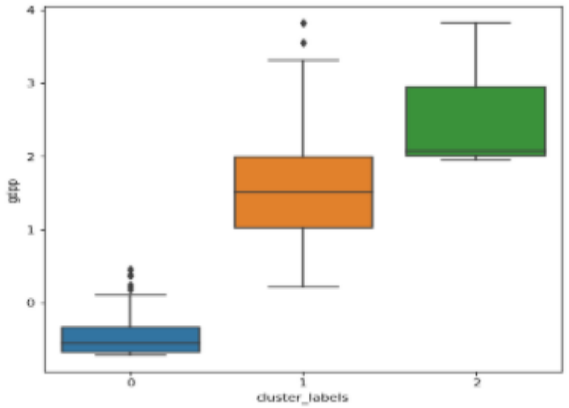


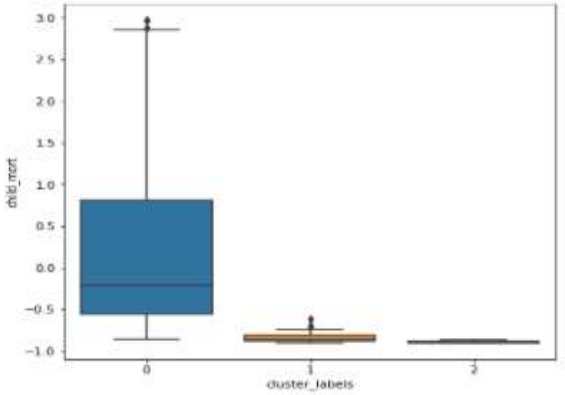

INSIGHT:
Here we can make out 3 distinct and dissimilar Clusters are obtained from the above dendrogram

# HIERARCHIAL CLUSTERING

- We re-ran the process of assigning Cluster labels and plotting a box plot with GDP per person, income and mortality in the y axis and Cluster_labels in the x axis




INSIGHTS:

We get a list of similar Cluster of countries

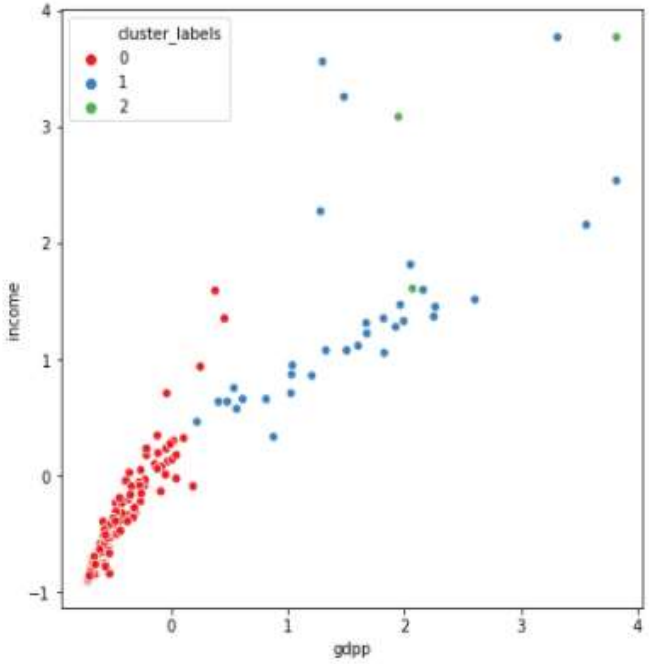
Cluster 0: Under developed Countries-Countries with low income and low GDP per person and high child mortality

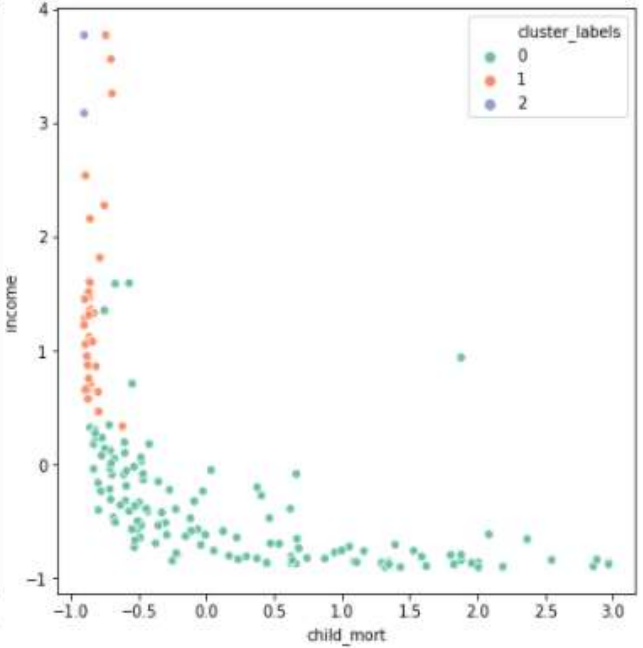
Cluster 1: Developing Countries- Countries with average income and average GDP per person and quite low child mortality rate

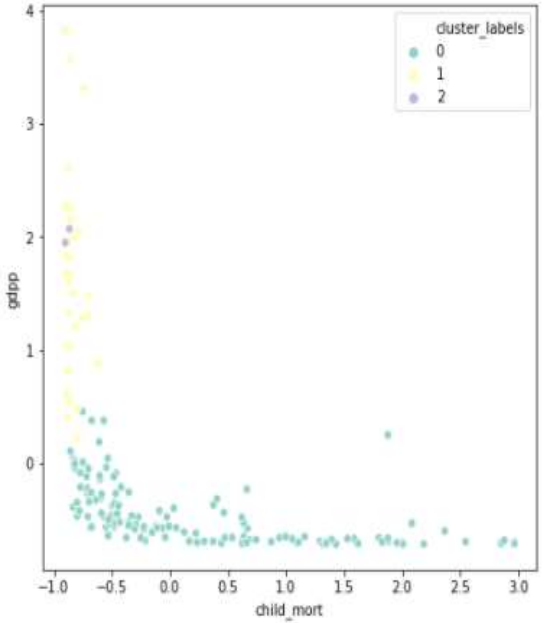
Cluster 2: Developed Countries- Countries with high income and high GDP per person and low child mortality rate

# VISUALIZATION COMPARING TWO PARAMETERS

- Plotting scatter plots by comparing **GDP and income, child mortality and income, and child mortality and GDP**




# LIST OF COUNTRIES

- 10 Countries we sorted on the basis of **low income, low GDP per person and high child mortality from Cluster 0**
- $ 10 million can be distributed among these Underdeveloped Nations by the NGO

1. Post conducting K mean analysis and Hierachical clustering we can see the list of countries are same in both cases.
2. The following 10 nations needs the financial aid from the NGO:

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

# Thank You!