Clustering Presentation

By,

Gevi Ackshaya

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, HELP 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.

Step 1 – Reading the data

The initial processes are performed as the following:

- Importing the necessary libraries
- Reading the data set
- Checking the variables data types
- Checking the shape
- Checking the uniqueness of the variables
- Checking for the null values

Step 2 – Performing EDA

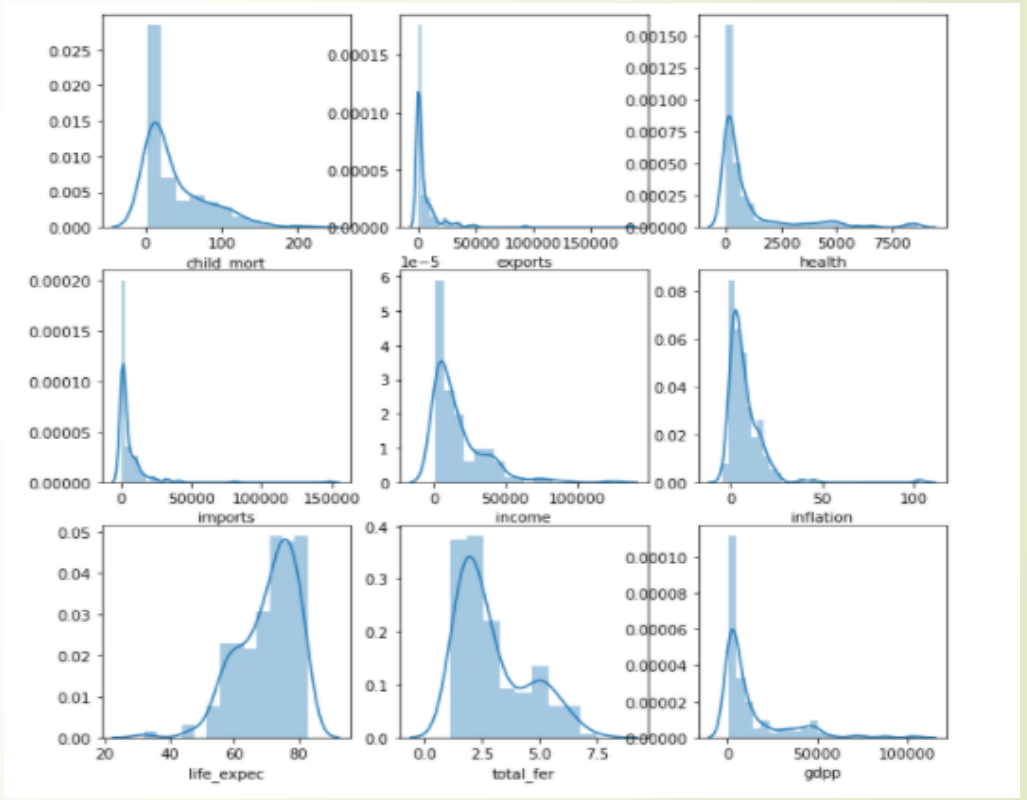
The EDA is performed in the given data-set

- Univariate Analysis
- Bivariate Analysis

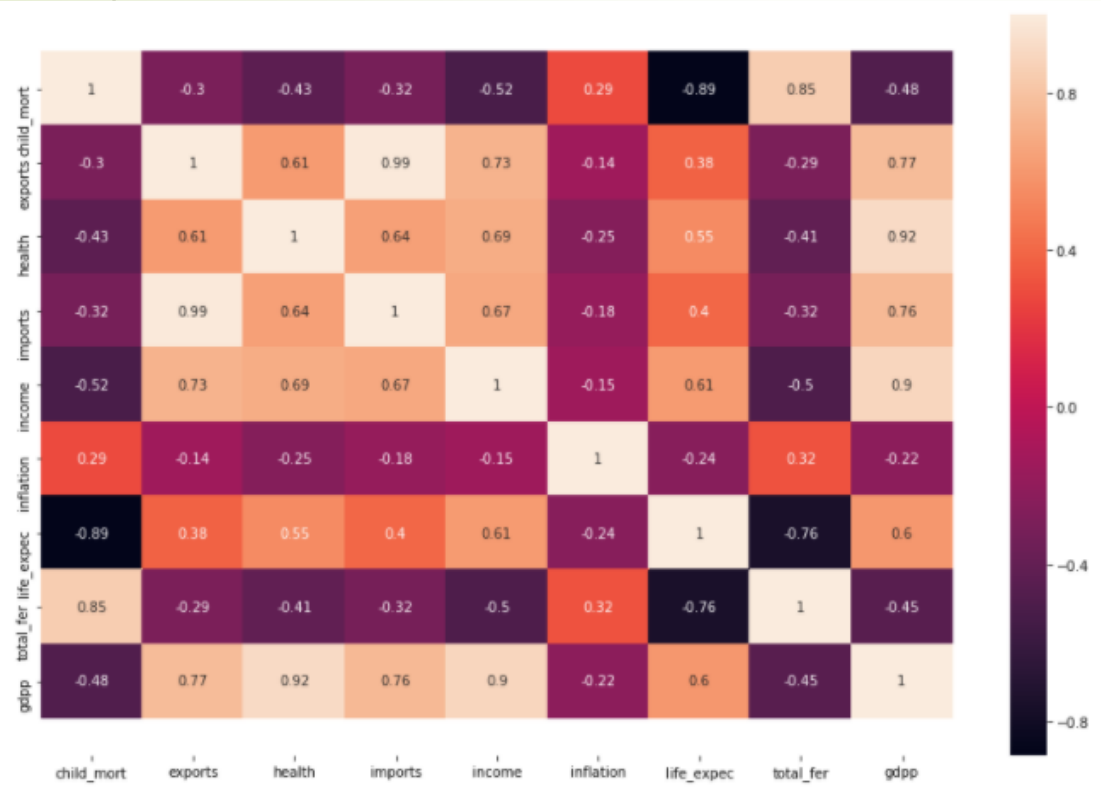
1. After the analysis the outliers are treated either by capping or removing it.
2. The analysis is done to find any patterns in the dataset.
3. The correlation between the variables are analysed.

Univariate Analysis

- From the above analysis it is evident that all the variables follow the normal distribution.
- These are distributions with variations and its not a smooth analysis.



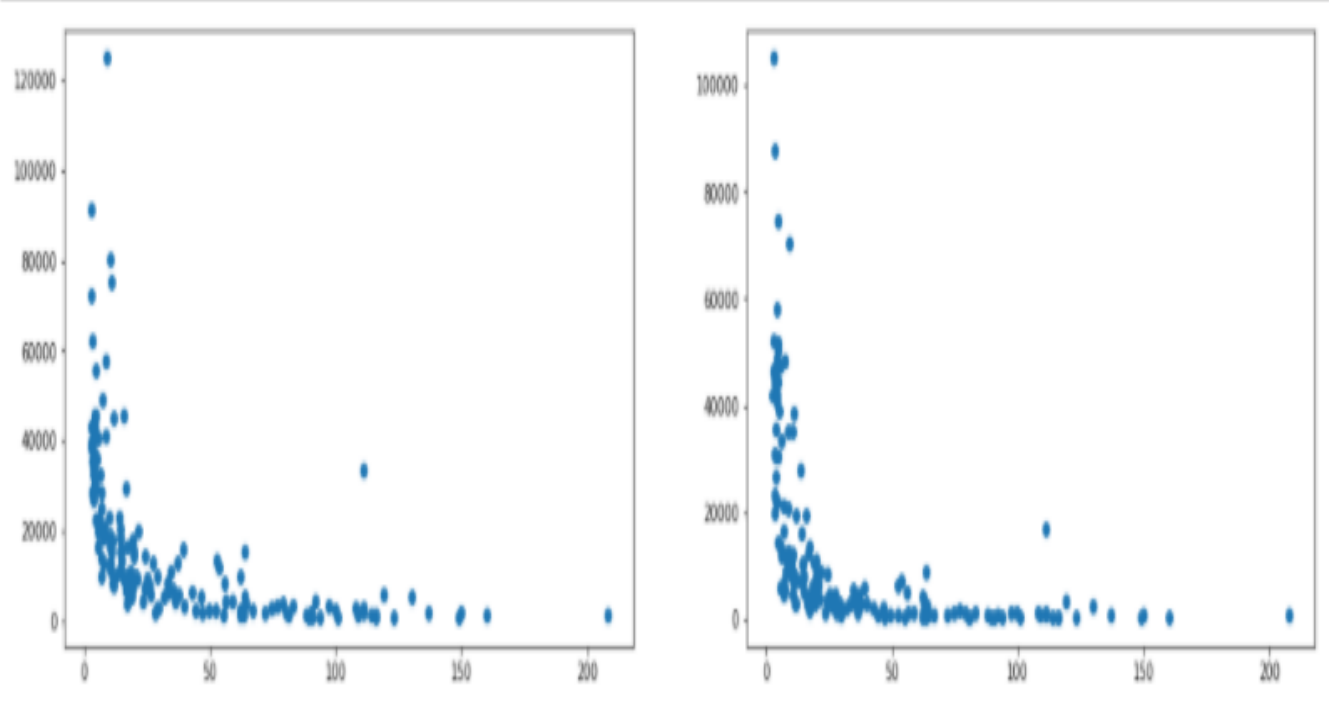
Correlation Matrix



From the above correlation matrix its shows that,

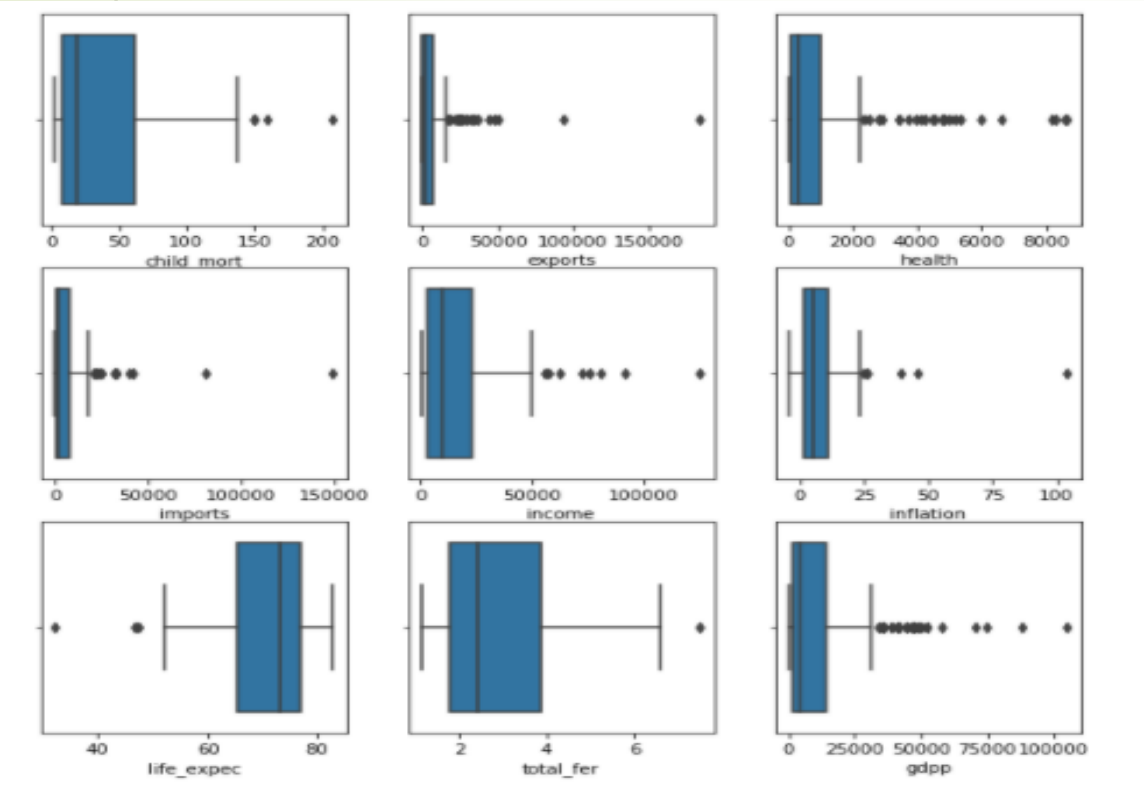
- 1. the gdpp variable is highly correlated to income and health.
- 2. the child_mort is highly correlated to total_fer and inversely correlated to life_expec
- 3. the income variable is highly correlated to health and gdpp and is inversely correlated to inflation

Bivariate Analysis



- The bivariate analysis was done between the variables child_mort, income and gdpp.
- The analysis shows that there are a few outliers present in the plot.
- The analysis shows that the points are left oriented.

Outlier Analysis



- The boxplot shows that there are few outliers present in the variables that needs to be treated either by removing/capping the values.

Capping the outliers

```
# capping the upper range of outliers in gdpp
q4 = data['gdpp'].quantile(0.99)
data['gdpp'][data['gdpp']>= q4] = q4
```

```
# capping the upper range of outliers in income
q4 = data['income'].quantile(0.95)
data['income'][data['income']>= q4] = q4
```

```
# capping the upper range of outliers in inflation
q4 = data['inflation'].quantile(0.95)
data['inflation'][data['inflation']>= q4] = q4
```

```
# capping the upper range of outliers in health
q4 = data['health'].quantile(0.95)
data['health'][data['health']>= q4] = q4
```

- The outliers are treated by capping the upper range of outliers in the variables

Scaling and Hopkins Score

```
# hopkins score
hopkins(data)
```

```
0.9701874368986969
```

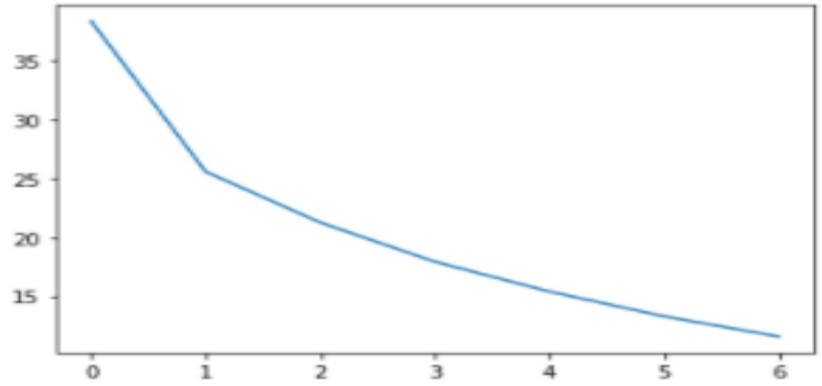
```
# scaling the data before modelling
from sklearn.preprocessing import MinMaxScaler
scaler = MinMaxScaler()

# fit_transform
data_scaled = scaler.fit_transform(data)
data_scaled.shape
```

```
(167, 9)
```

- the Hopkins score is calculated
- The values of the variables are scaled using the Min-Max Scaler

K-Means Cluestering

```
[<matplotlib.lines.Line2D at 0xae12ede808>]
```




```
For n_clusters=2, the silhouette score is 0.48207528073614514
For n_clusters=3, the silhouette score is 0.39323527053787205
For n_clusters=4, the silhouette score is 0.3442514921071971
For n_clusters=5, the silhouette score is 0.35915292602917986
For n_clusters=6, the silhouette score is 0.331917171557936
For n_clusters=7, the silhouette score is 0.34944825523074213
For n_clusters=8, the silhouette score is 0.3448427137849402
```

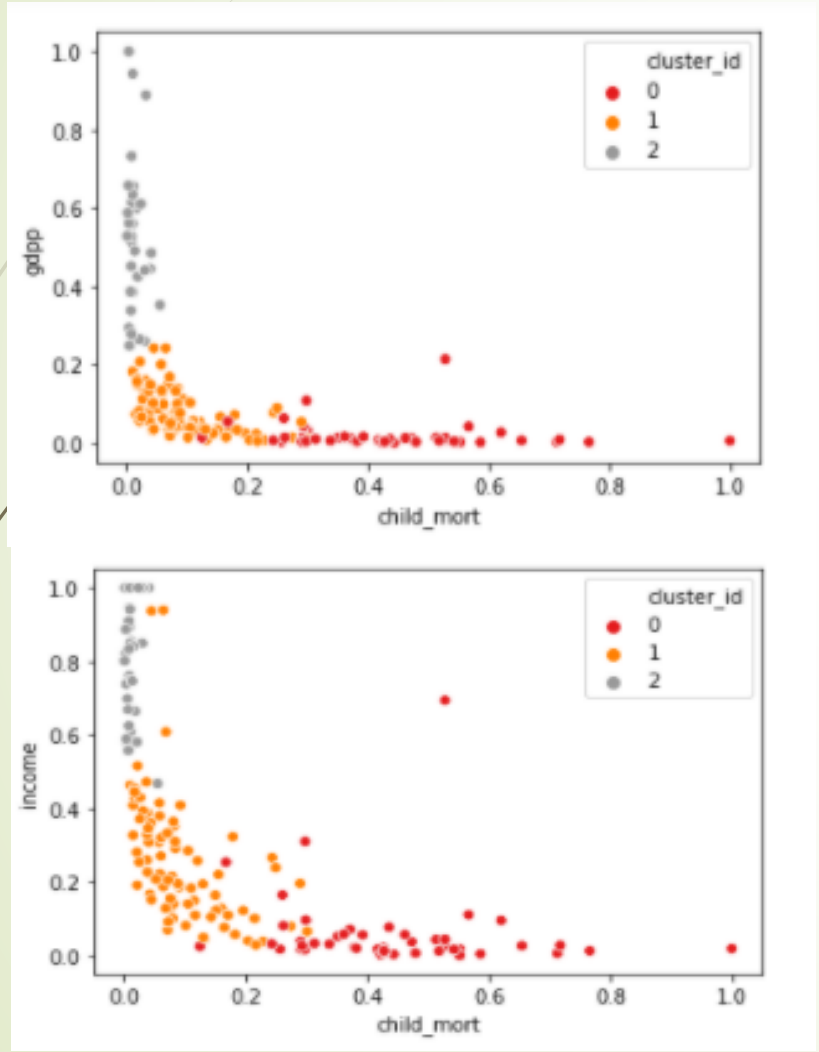
- Using the Elbow curve and the Silhouette scores the ideal k value is determined as **3**

Steps Followed in K-Means

Once the number of cluester are decided

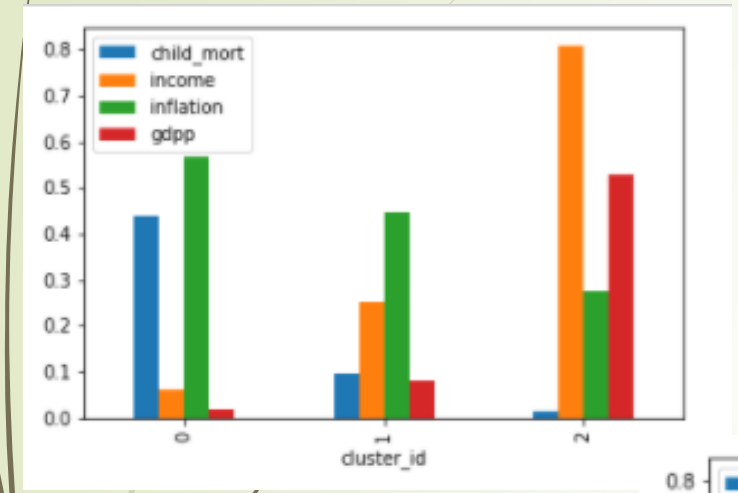
- The cluster ID's are generated
- The cluster_id is attached to the newly created data frame
- The cluster profiling is done and the visualizations are done
- Based on the visualizations the cluster that needs AID is analysed and then a separate data frame is created with only the cluster that needs AID

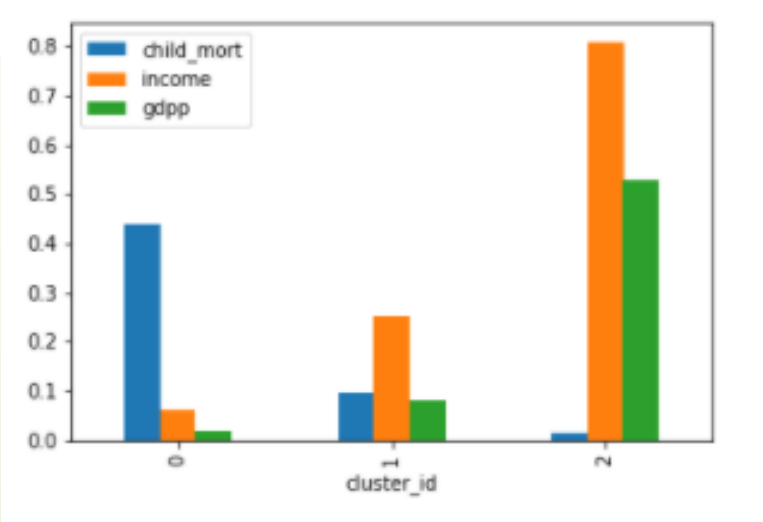
Cluster Profiling



- From the plot it is evident that the cluster_id-0 is the one that needs to be given AID based on the analysis of child_mort with the GDPP and income .
- The gdpp and income values of the other two clusters are pretty much high than cluster_id = 0

Cluster Profiling





- The bar graph analysis show that the cluster=0 has the highest child mortality rate with the least income and gdpp rates than the other two clusters.
- Hence its evident that the cluster=0 is the one that needs to be given AID.

Result => K-Means Algorithm

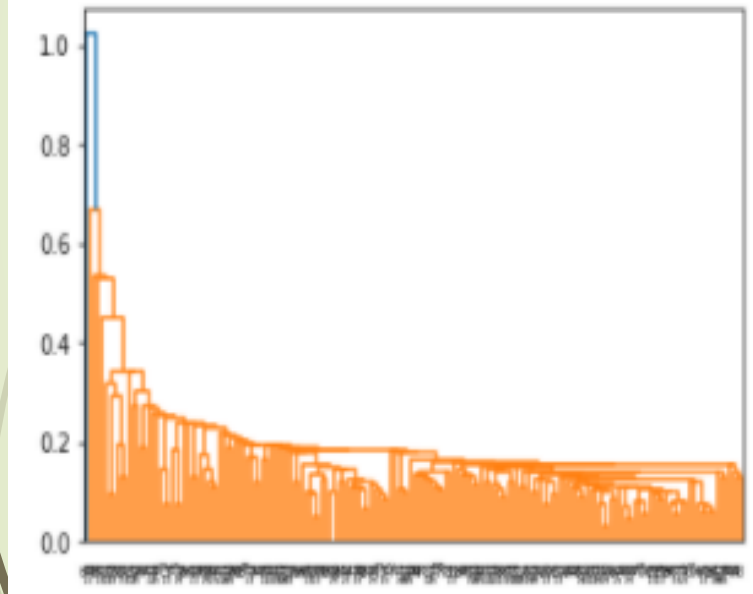
According to the K-Means Algorithm,

These are the countries that needs to be given AID:

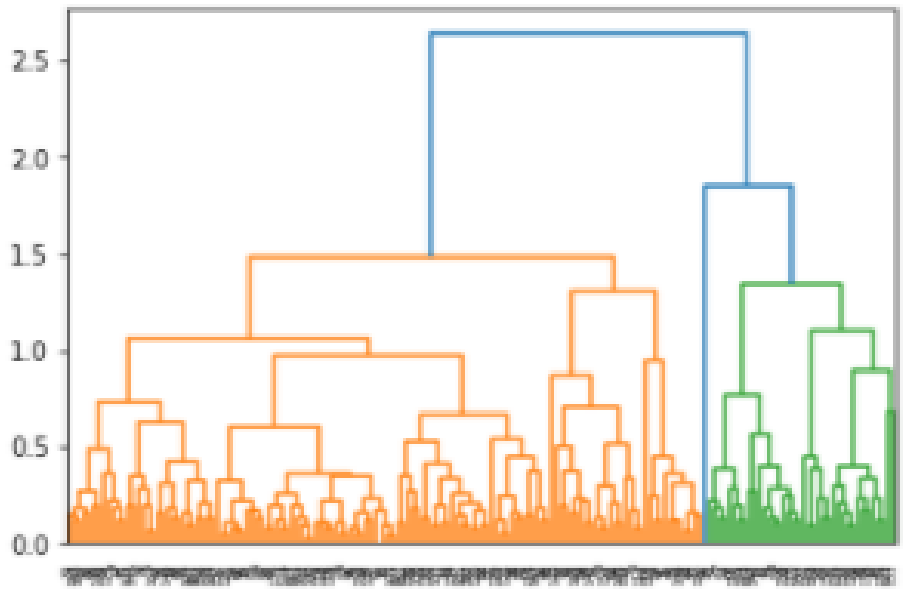
- Afghanistan, Angola, Benin, Burkina Faso, Burundi, Cameroon, Central African Republic, Chad, Comoros, Congo, Dem. Rep., Congo, Rep., Cote d'Ivoire, Equatorial Guinea, Eritrea, Gabon, Gambia, Ghana, Guinea, Guinea-Bissau, Haiti, Iraq, Kenya, Kiribati, Lao, Lesotho, Liberia, Madagascar, Malawi, Mali,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

Hierarchical Clustering

```
# single linkage
mergings = linkage(data_2, method="single", metric='euclidean')
dendrogram(mergings)
plt.show()
```



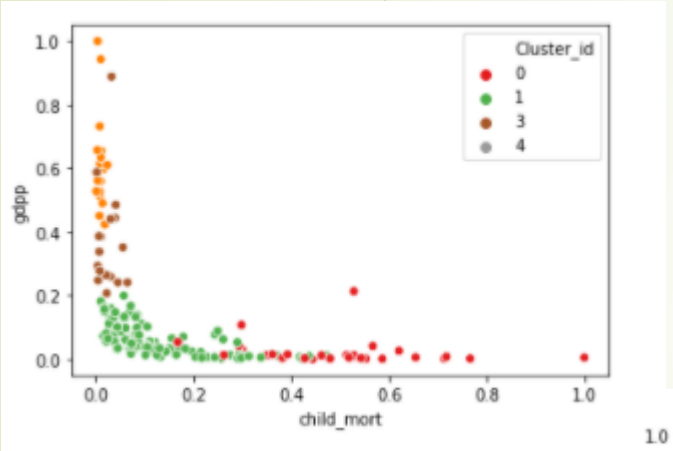

```
# complete linkage
mergings = linkage(data_2, method="complete", metric='euclidean')
dendrogram(mergings)
plt.show()
```

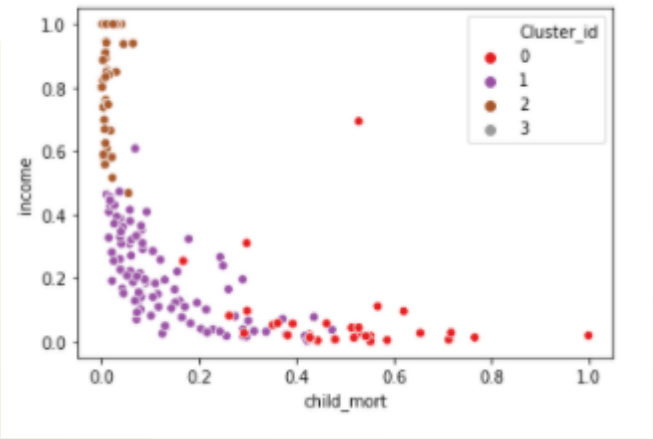


Steps in Hierarchical Clustering

- After the clusters are visualized based on the single linkage and complete linkage method, the number of cluster is decided as **4.**
- The cluster ids are generated based on the hierarchal clustering.
- The cluster_ids are added as a column to the new data frame.
- The cluster visualization is done and then the cluster that needs AID is identified

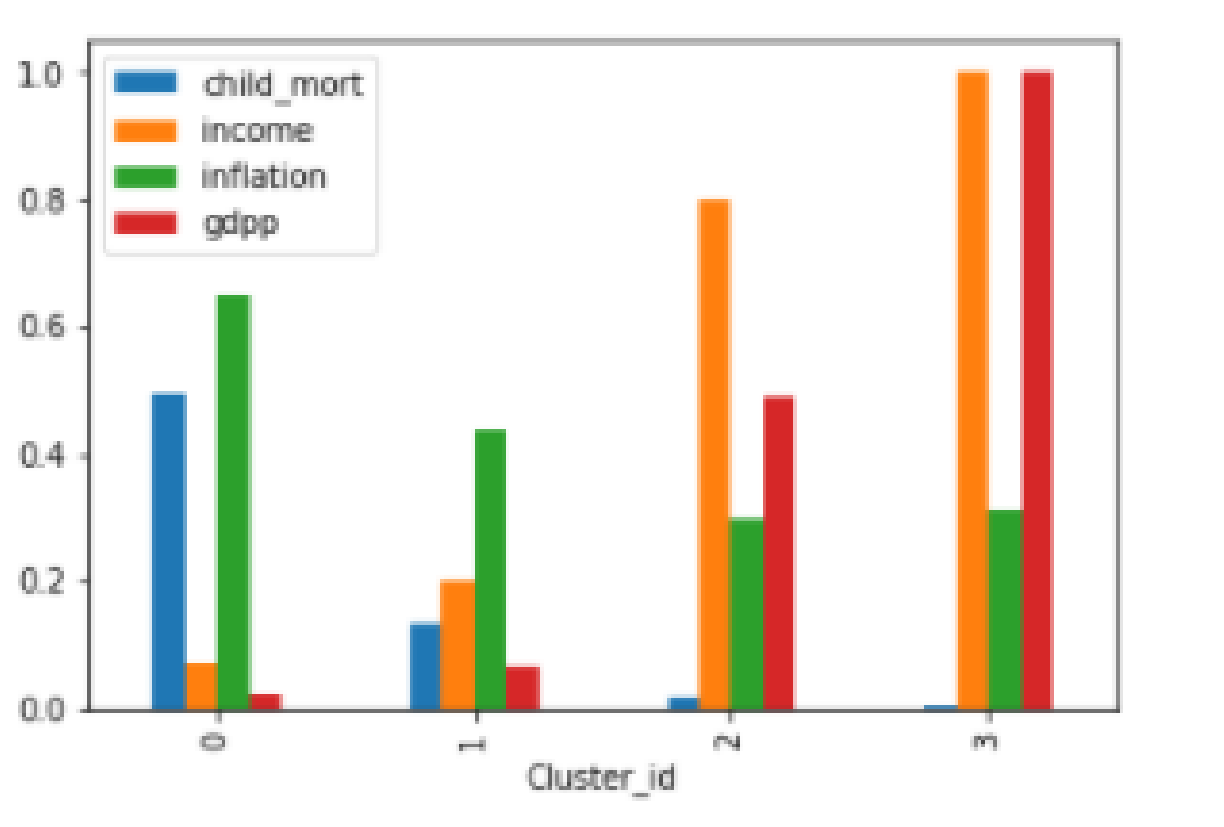
Cluster Profiling





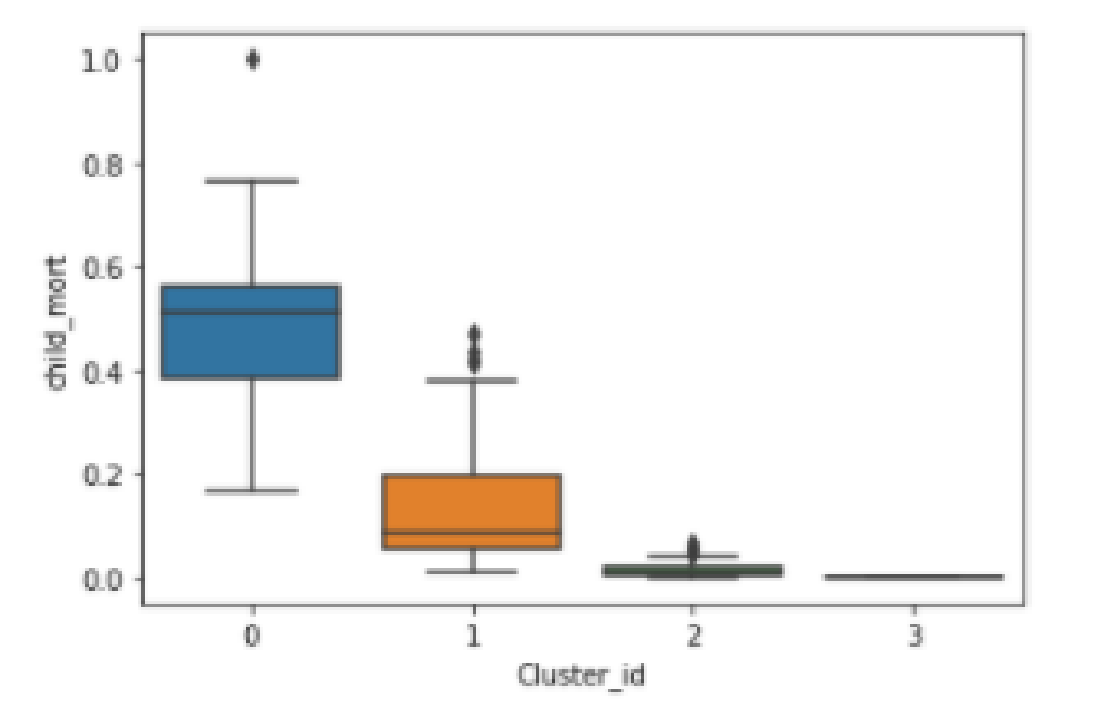
- From the plot it is evident that the cluster_id 0 is the one that needs to be given AID based on the analysis of child_mort with the income and gdpp.
- The income and gdpp values of the other two clusters are pretty much high than cluster_id = 0

Cluster Profiling



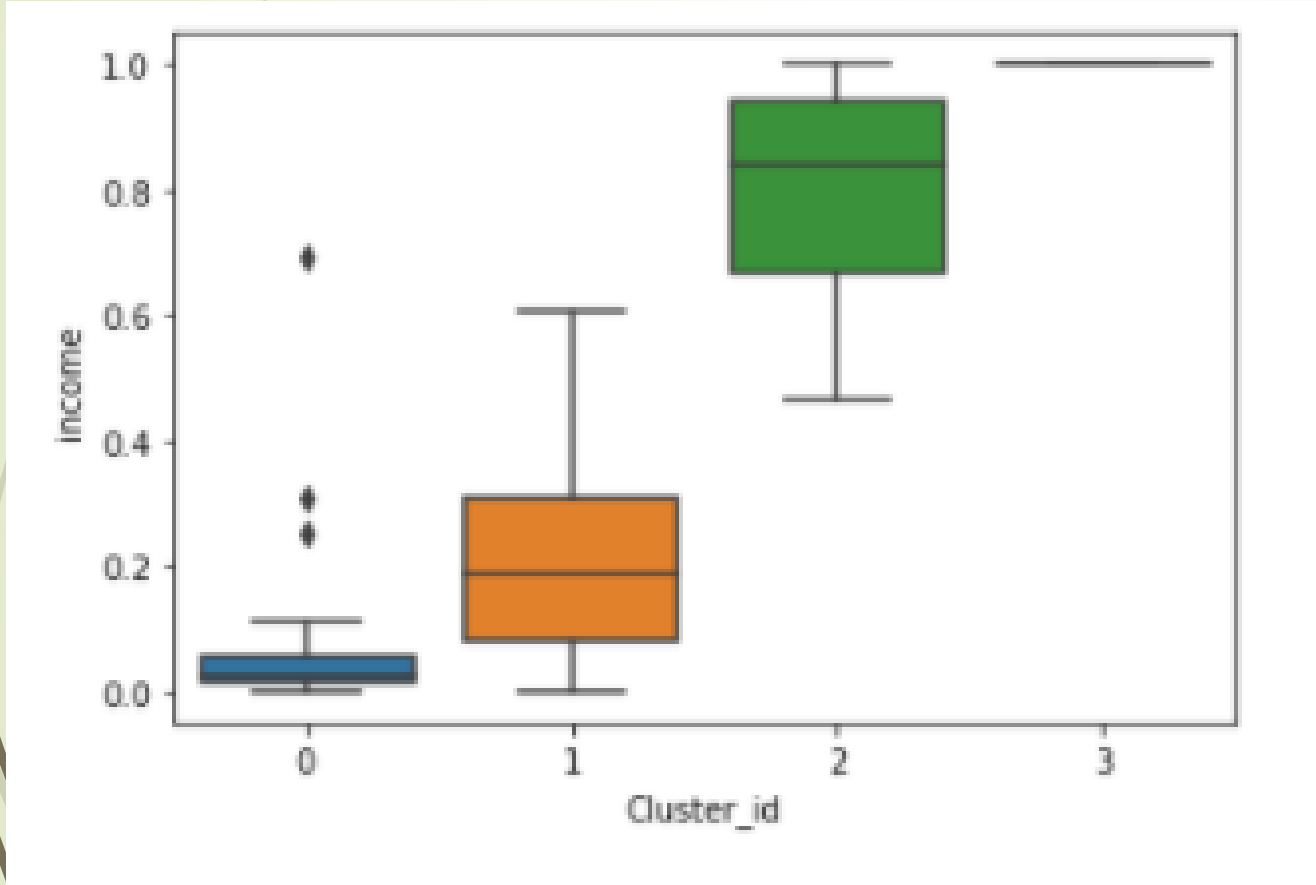
- The above bar graph analysis show that the cluster=0 has the highest child mortality rate with the least income and gdpp rates than the other two clusters.
- Hence its evident that the cluster=0 is the one that needs to be given AID.

Visualization of variables



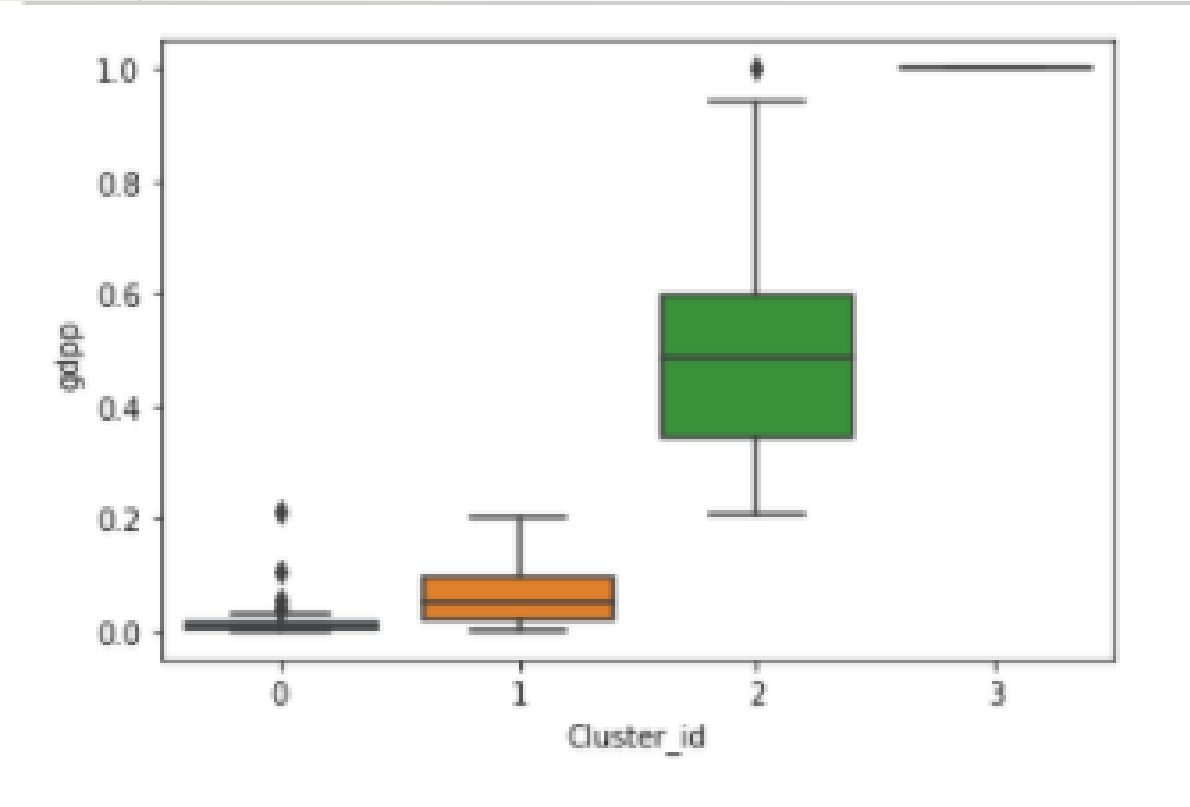
- The boxplot very clearly that the child mortality rate is the highest for cluster_id = 0

Visualization



- The boxplot shows the income value of the cluster 0 is very less than any other clusters.

Visualization - GDPP



- The boxplot clearly mentions that the gdpp value of the cluster 0 is very less than any other clusters

Results – Hierarchical Clustering

According to the hierarchical clustering,

These are the countries that needs to be given AID:

Afghanistan, Angola, Benin, Burkina Faso, Burundi, Cameroon, Central African Republic, Chad, Congo, Dem. Rep., Congo, Rep., Cote d'Ivoire, Equatorial Guinea, Gabon, Ghana, Guinea, Guinea-Bissau, Haiti, Iraq, Malawi, Mali, Mauritania, Mozambique, Niger, Nigeria, Sierra Leone, Sudan, Timor-Leste, Togo, Uganda, Yemen, Zambia

Analysis Decision

According to the analysis of the K-Means and hierarchical clustering, it is good to follow Hierarchical clustering because,

- The number of clusters formed is more
- The number of countries mentioned in the cluster that needs AID is less than that of K-Means
- On the basis of business and technical perspective, less the no. of countries that needs to given AID higher would be the quality and AID's provided for it.
- Nearly 16 countries were less in the cluster that was provided by hierarchical clustering.

Countries that needs AID

- Afghanistan, Angola, Benin, Burkina Faso, Burundi, Cameroon, Central African Republic, Chad, Congo, Dem. Rep., Congo, Rep., Cote d'Ivoire, Equatorial Guinea, Gabon, Ghana, Guinea, Guinea-Bissau, Haiti, Iraq, Malawi, Mali, Mauritania, Mozambique, Niger, Nigeria, Sierra Leone, Sudan, Timor-Leste, Togo, Uganda, Yemen, Zambia

Top 5 countries

The top 5 countries that needs immediate AID by sorting the cluster with high mortality rate and less income, gdpp rates are,

- Haiti
- Sierra Leone
- Chad
- Central African Republic
- Mali