Analysis of the Clustering Assignment (Unsupervised Learning)

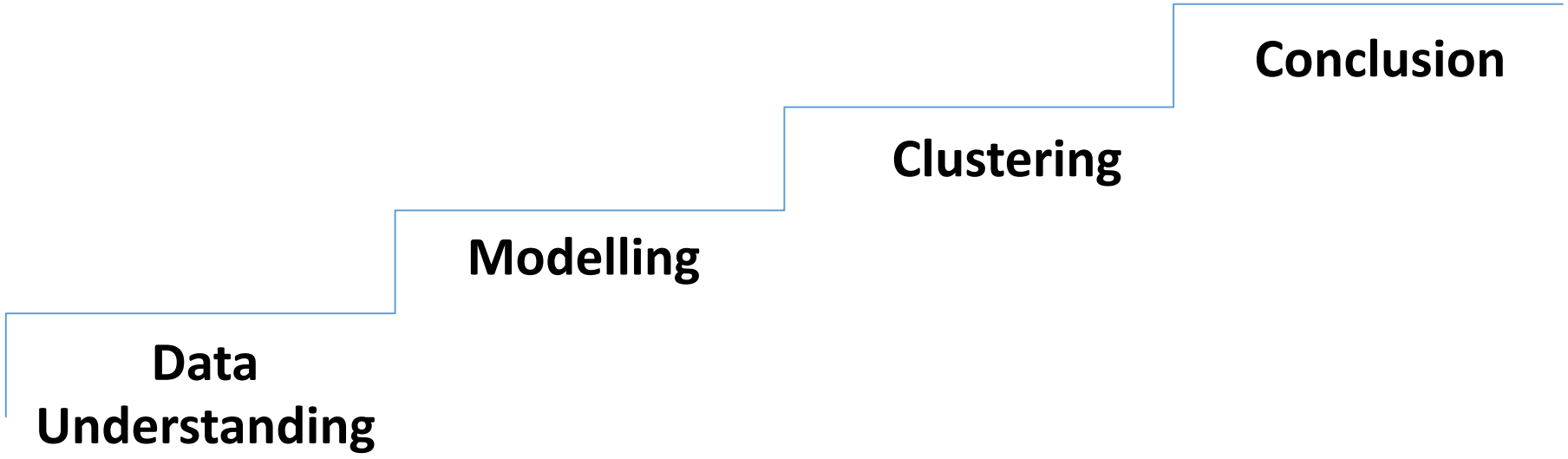
Dataset :

Countries Data

Problem Statement :

To identify the top 5 countries which are in direst need of aid so that the HELP International (which is an international humanitarian NGO that is committed to fighting poverty) can provide the people of backward countries with basic amenities and relief during the time of disasters and natural calamities by providing them funds.

All the Steps Involved in the Analysis

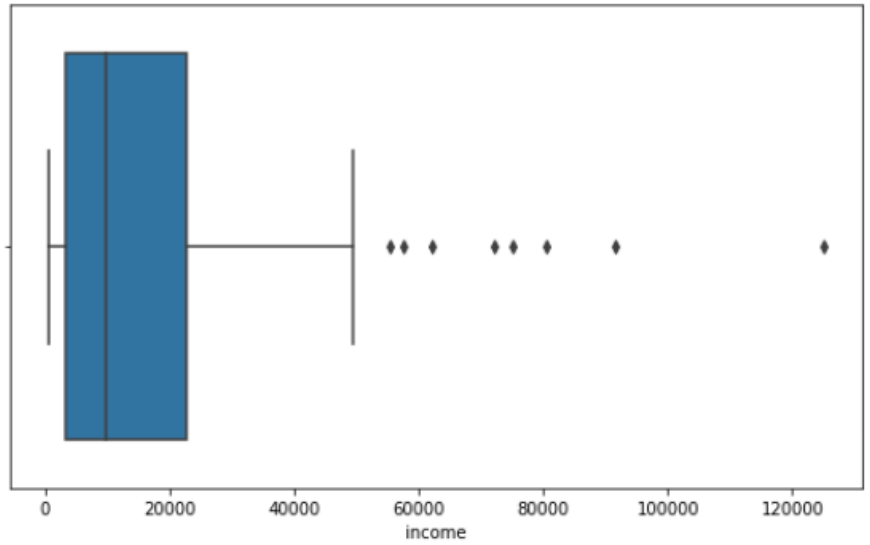


Data Understanding – Data Stats

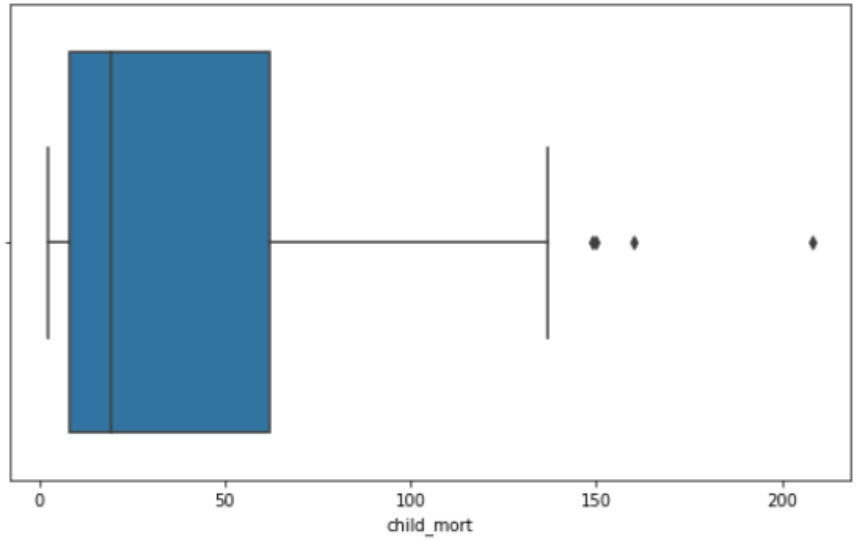
- There are 167 rows and 10 columns present in the dataset.
- None of the columns had missing values.
- Converted exports, health and imports from percentage value of GDP per capita to their actual value.

Data Understanding – Checking for Outliers

- Column Name – **income**
- Description : Net income per person
- The plot shows that they are some outliers in the higher range income values.

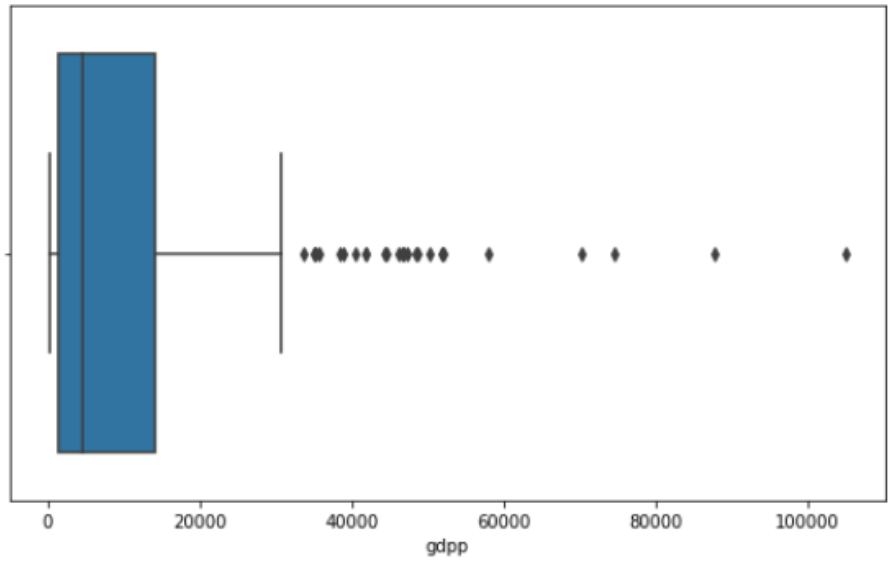


- Column Name – **child_mort**
- Description : Death of children under 5 years of age per 1000 live births
- The plot indicates that there are some countries where they are very high child deaths.

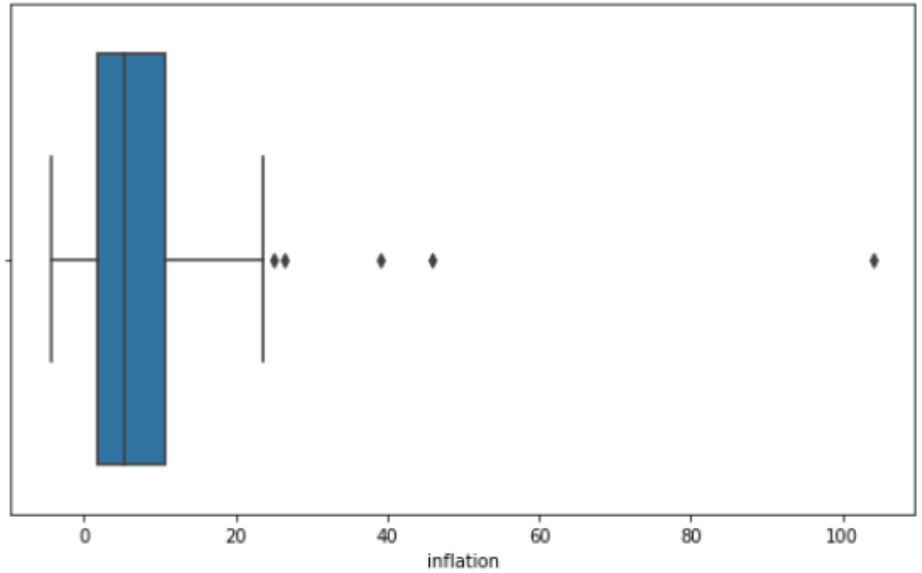


Data Understanding – Checking for Outliers

- Column Name – **gdpp**
- Description : The GDP per capita. Calculated as the Total GDP divided by the total population.
- The plot shows that there are many countries where the gdpp value is very high

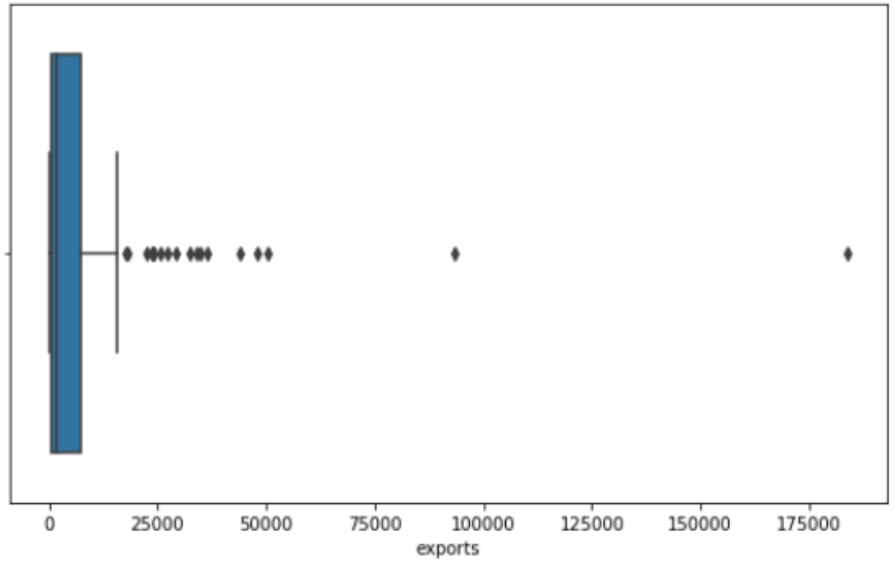


- Column Name – **inflation**
- Description : The measurement of the annual growth rate of the Total GDP.
- The plot indicates there is one country where the inflation value is too high(outlier needs to be treated)

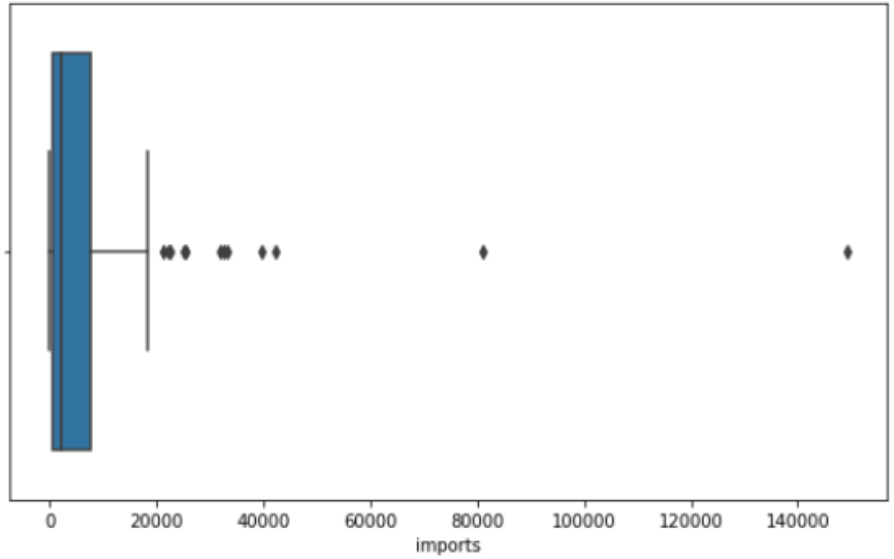


Data Understanding – Checking for Outliers

- Column Name – **exports**
- Description : Exports of goods and services per capita. Given as %age of the GDP per capita
- The plot shows that there is one country having exports outlier value which is very high.



- Column Name – **imports**
- Description : Imports of goods and services per capita. Given as %age of the GDP per capita
- The plot indicates there is extreme outlier having import value greater than 140000



Data Understanding – Outliers Treatment

Identifying Outliers with Skewness

- Explains the extent to which the data is normally distributed.
- Ideally, the skewness value should be between -1 and +1, and any major deviation from this range indicates the presence of extreme values.

- Table 1 – Indicates skewness before capping the outliers

- Table 2 – Indicates skewness after capping the outliers

TABLE 1

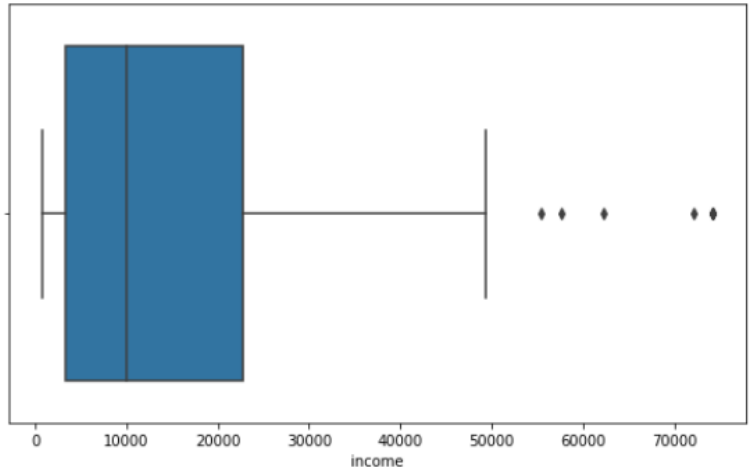
Column Names	Skew Value
income	2.23148
child_mort	1.450774
gdpp	2.218051
inflation	5.154049
exports	6.720171
health	2.526029
imports	6.6185

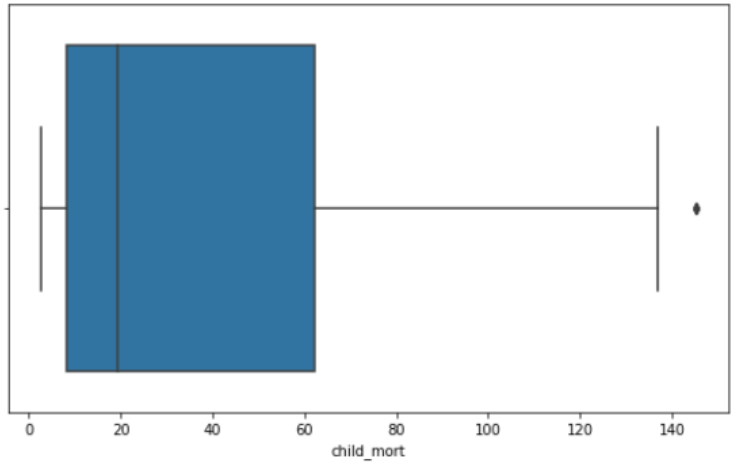
TABLE 2

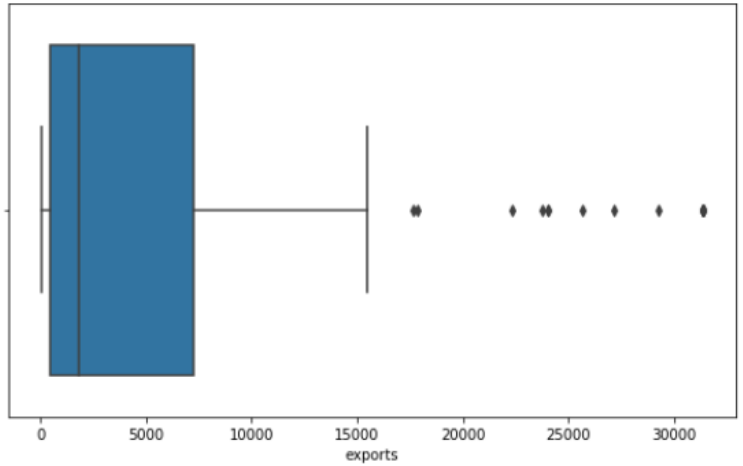
Column Names	Skew Value
income	1.527598
child_mort	1.212276
gdpp	1.702615
inflation	1.061468
exports	1.937935
health	2.526029
imports	1.698238

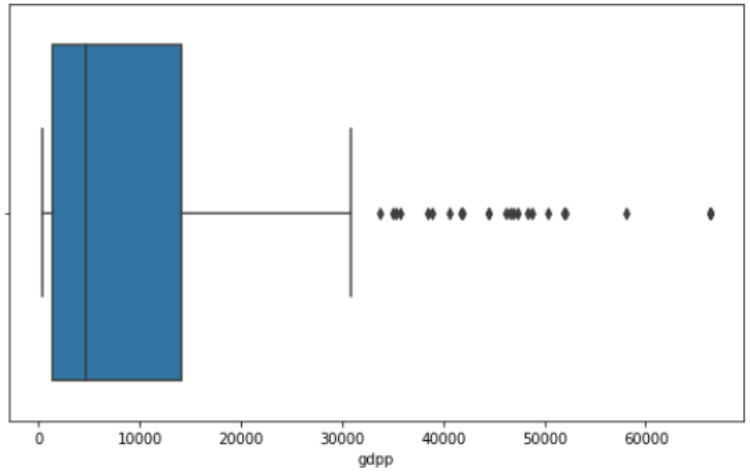
Data Understanding – Outliers Treatment

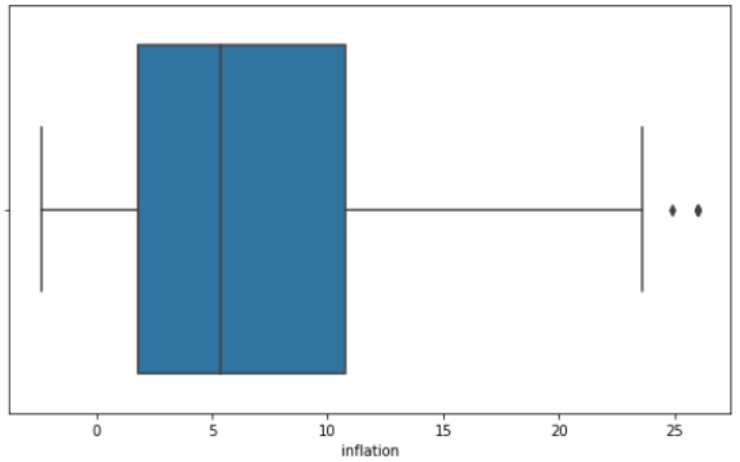
Plots after capping the outliers

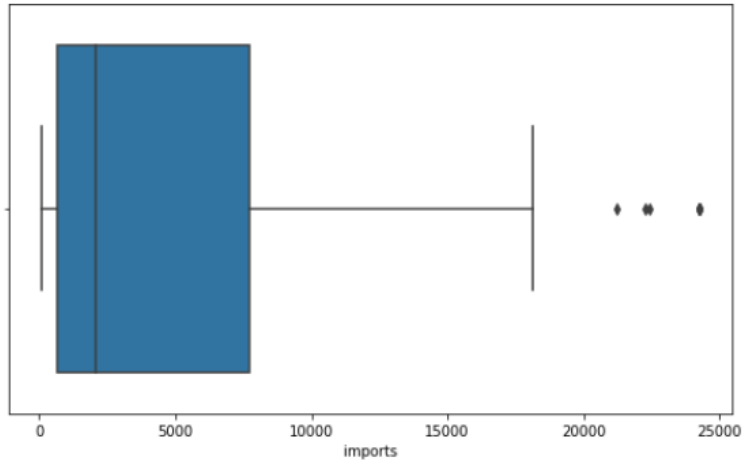












Modelling – Rescaling Columns

- Rescaling the columns Using Standard Scalar , so that all the columns are scaled before clustering.

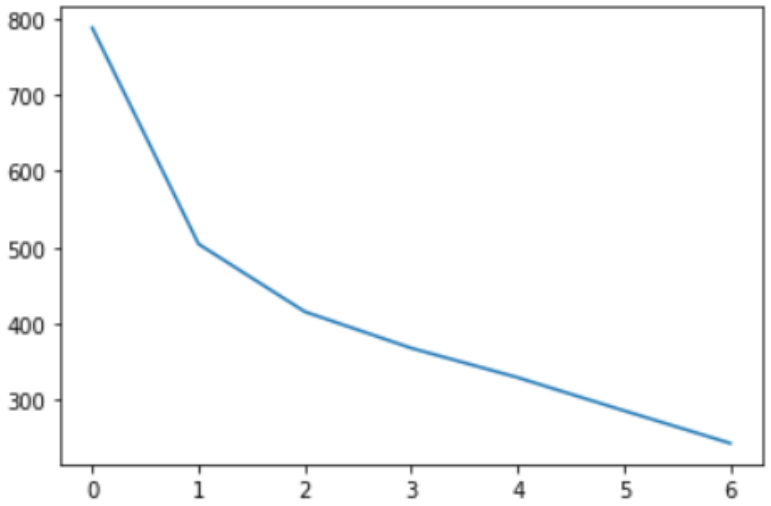
- Using **Hopkins Check** to evaluate the data is feasible for clustering or not (i.e. checking cluster tendency).

- The dataset has obtained a Hopkins Check value of **87.4 %** (Higher values indicates it has a high tendency form clusters)

Clustering – K Means Clustering

Finding the Optimal Number of Clusters

- **SSD/Elbow –curve**



- **Silhouette Analysis**

```
For n_clusters=2, the silhouette score is 0.4722260243921151
For n_clusters=3, the silhouette score is 0.40123143009349704
For n_clusters=4, the silhouette score is 0.3434886413731235
For n_clusters=5, the silhouette score is 0.3279894506610432
For n_clusters=6, the silhouette score is 0.29329131153768906
For n_clusters=7, the silhouette score is 0.33309347049050564
For n_clusters=8, the silhouette score is 0.3002326380612719
```

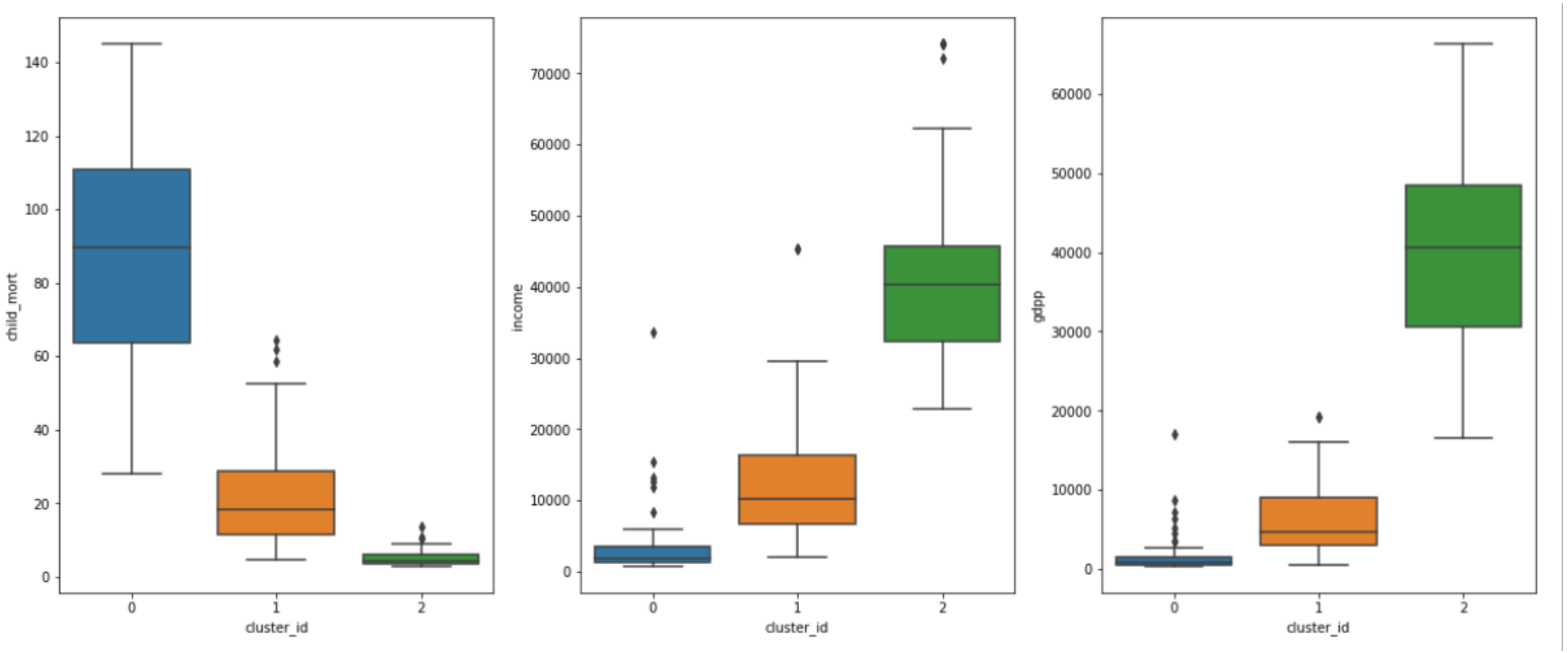
Conclusion

Based on the above methods, selecting cluster value(k) as 3

Clustering – K Means Clustering

Performing K means (k=3) clustering and plotting the visuals

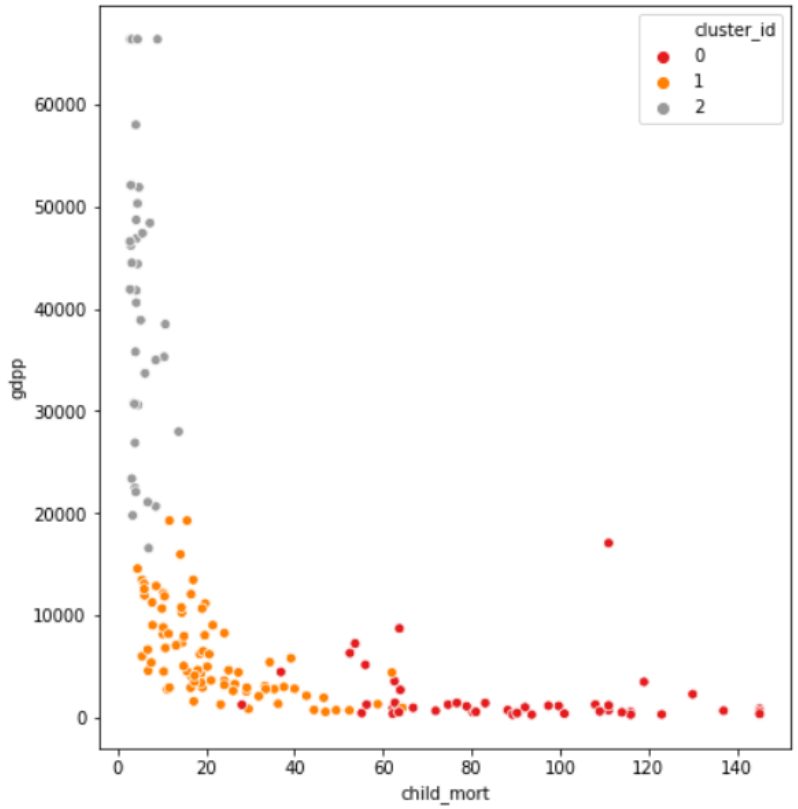
- Using Box Plot

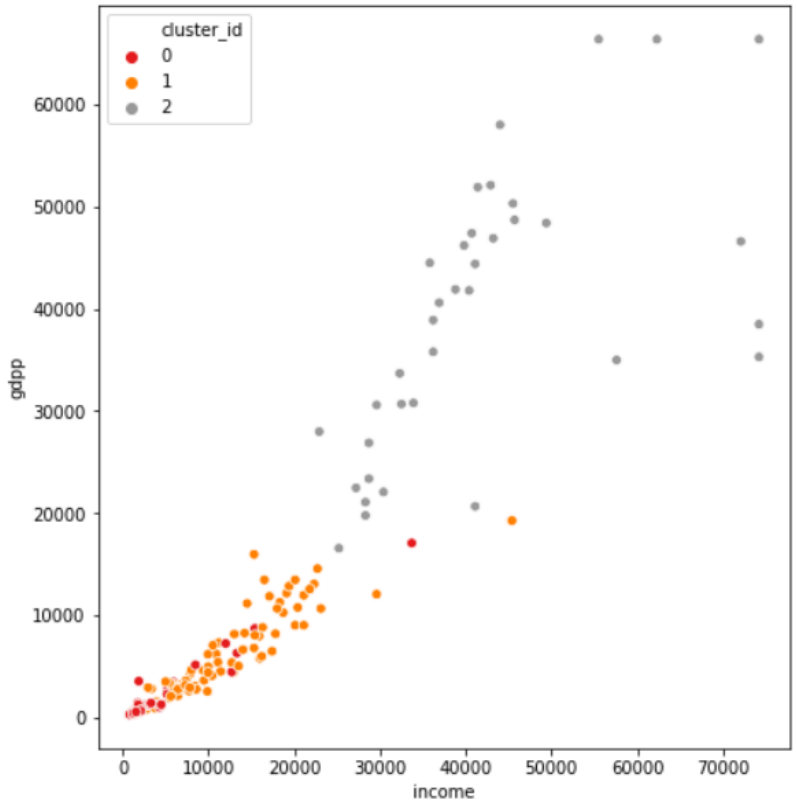


Clustering – K Means Clustering

Performing K means (k=3) clustering and plotting the visuals

- Using Scatter Plot



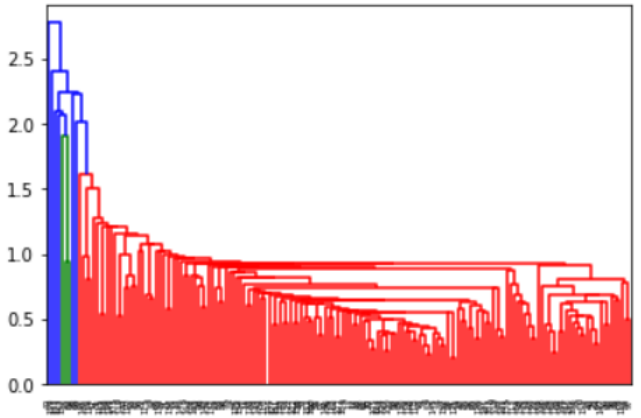


Conclusion

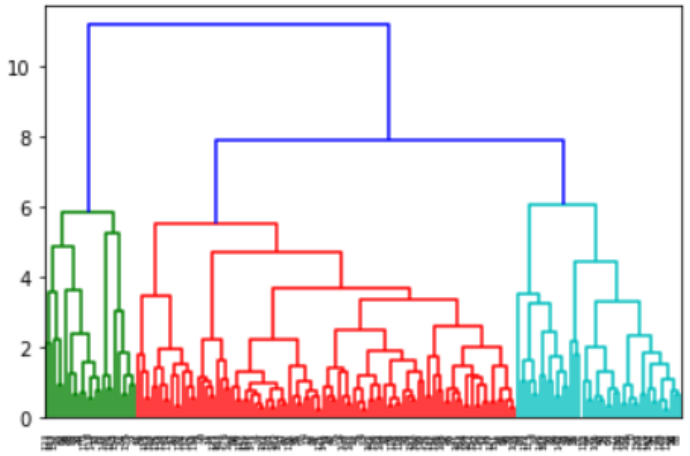
From the box and scatter plots graphs,we can conclude that **Cluster 0** is the one which is in dire need of aid, since they have lowest gdpp,lowest income and high child mortality rate

Clustering – Hierarchical Clustering

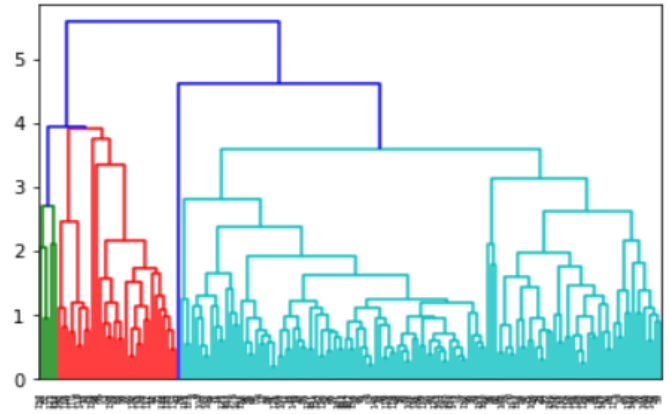
Plotting Dendogram using single Linkage



Plotting Dendogram using complete Linkage



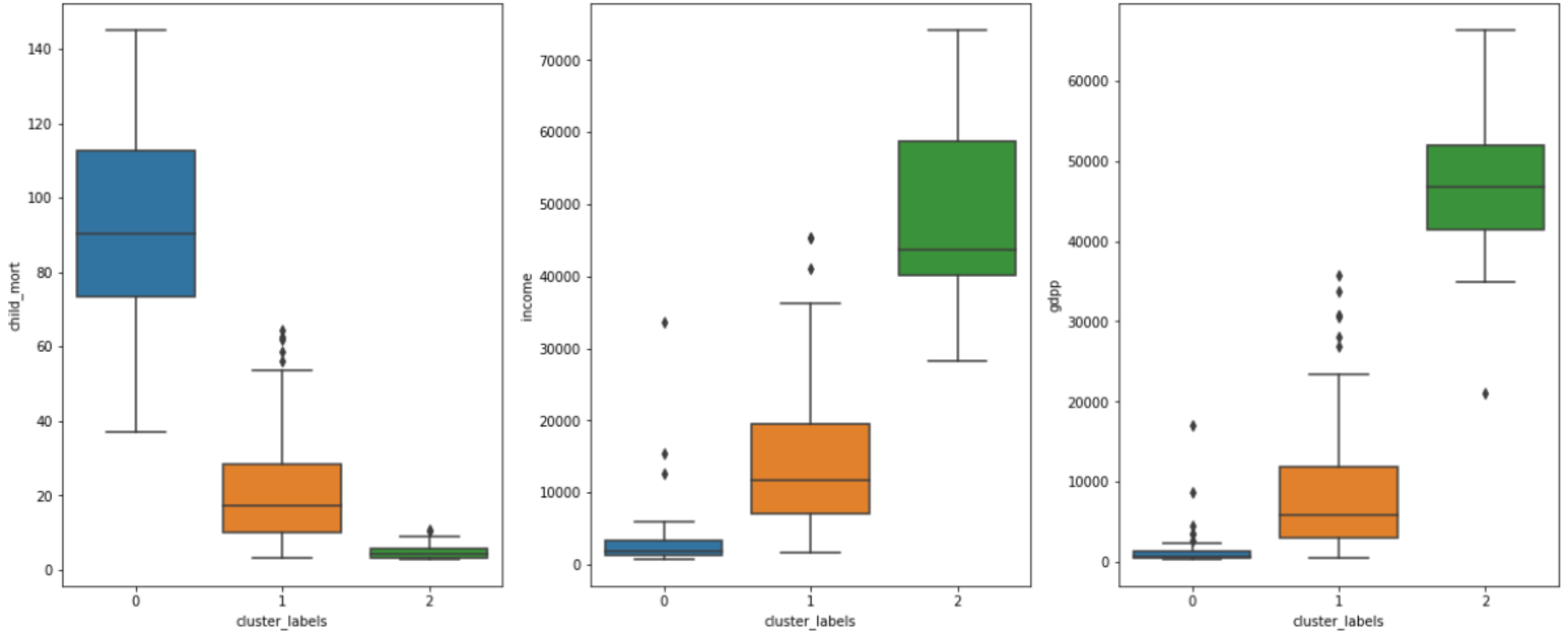
Plotting Dendogram using average Linkage



Clustering – Hierarchical Clustering

Cutting the dendogram at 3 based on complete linkage and plotting the visuals

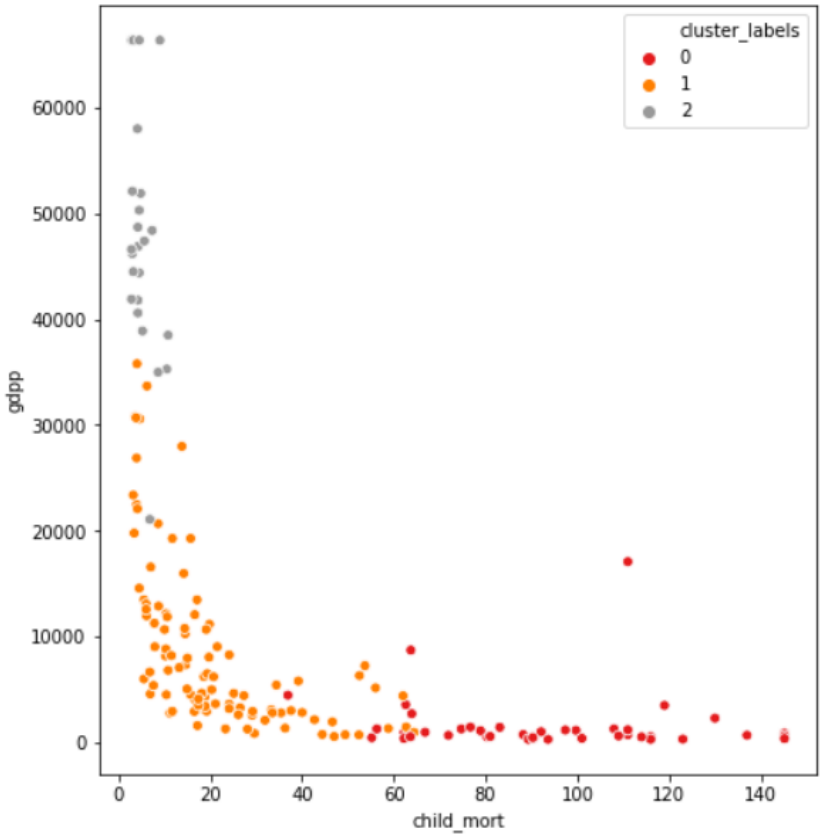
- Using box plot

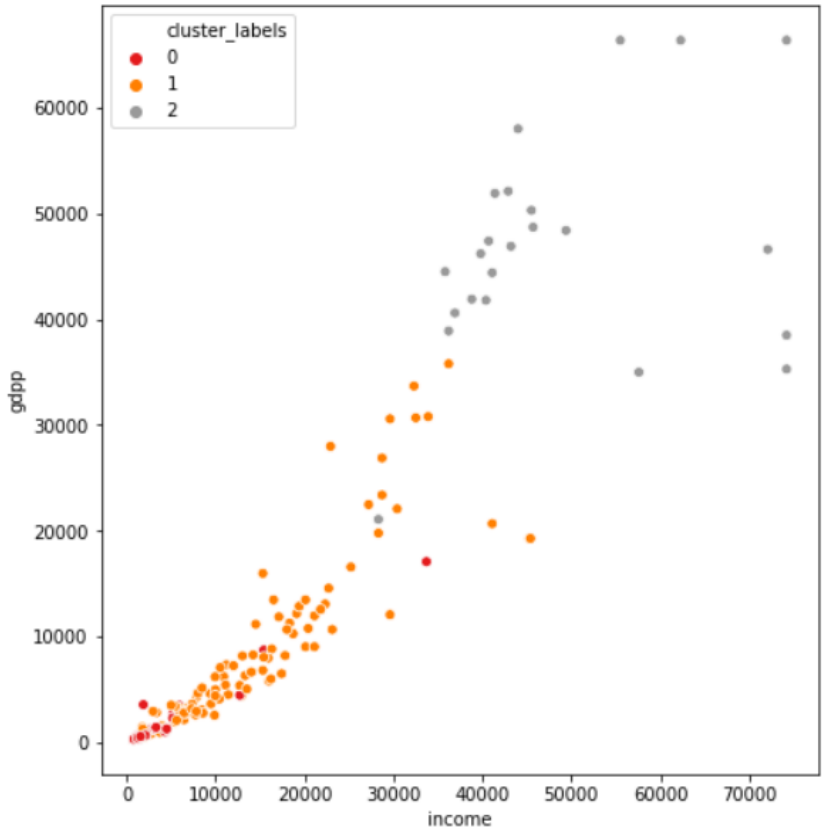


Clustering – Hierarchical Clustering

Cutting the dendogram at 3 based on complete linkage and plotting the visuals

- Using scatter plot





Conclusion

From the box and scatter plots graphs,we can conclude that **Cluster 0** is the one which is in dire need of aid, since they have lowest gdpp,lowest income and high child mortality rate

Conclusion

The top 5 countries which are in dire aid of need *(Based on lowest gdpp,lowest income and highest child mortality rate)*

Using K Means Clustering

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	cluster_id
88	Liberia	89.30	62.457000	38.5860	302.80200	742.24	5.47	60.8	5.02	331.62	0
26	Burundi	93.60	22.243716	26.7960	104.90964	764.00	12.30	57.7	6.26	331.62	0
37	Congo, Dem. Rep.	116.00	137.274000	26.4194	165.66400	742.24	20.80	57.5	6.54	334.00	0
112	Niger	123.00	77.256000	17.9568	170.86800	814.00	2.55	58.8	7.49	348.00	0
132	Sierra Leone	145.16	67.032000	52.2690	137.65500	1220.00	17.20	55.0	5.20	399.00	0

Using Hierarchical Clustering

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	cluster_id	cluster_labels
88	Liberia	89.30	62.457000	38.5860	302.80200	742.24	5.47	60.8	5.02	331.62	0	0
26	Burundi	93.60	22.243716	26.7960	104.90964	764.00	12.30	57.7	6.26	331.62	0	0
37	Congo, Dem. Rep.	116.00	137.274000	26.4194	165.66400	742.24	20.80	57.5	6.54	334.00	0	0
112	Niger	123.00	77.256000	17.9568	170.86800	814.00	2.55	58.8	7.49	348.00	0	0
132	Sierra Leone	145.16	67.032000	52.2690	137.65500	1220.00	17.20	55.0	5.20	399.00	0	0

Conclusion

K means and Hierarchical Clustering results in the same list of top 5 countries

Thank You