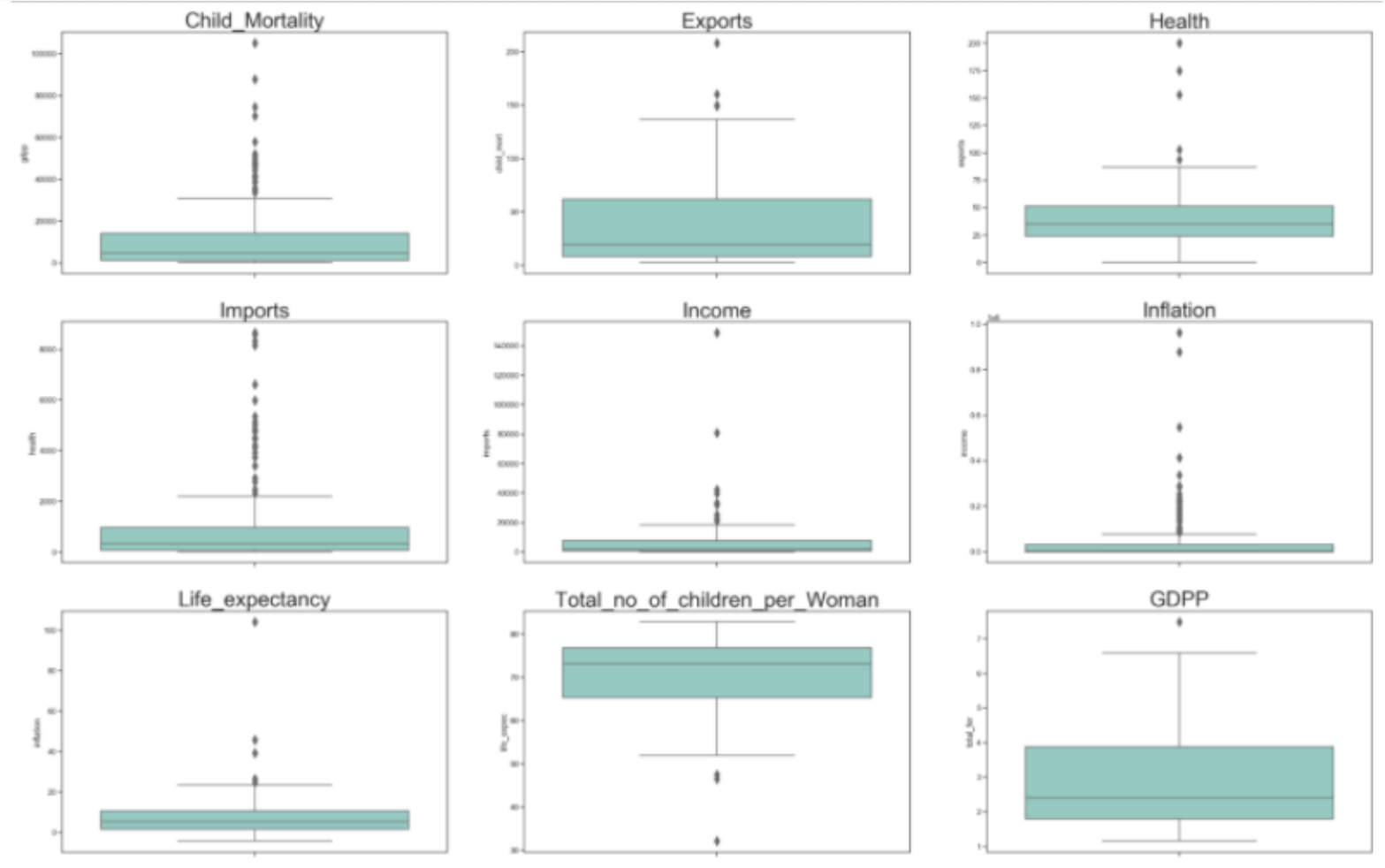
Clustering Assignment

-by Harish J

Data Visualization

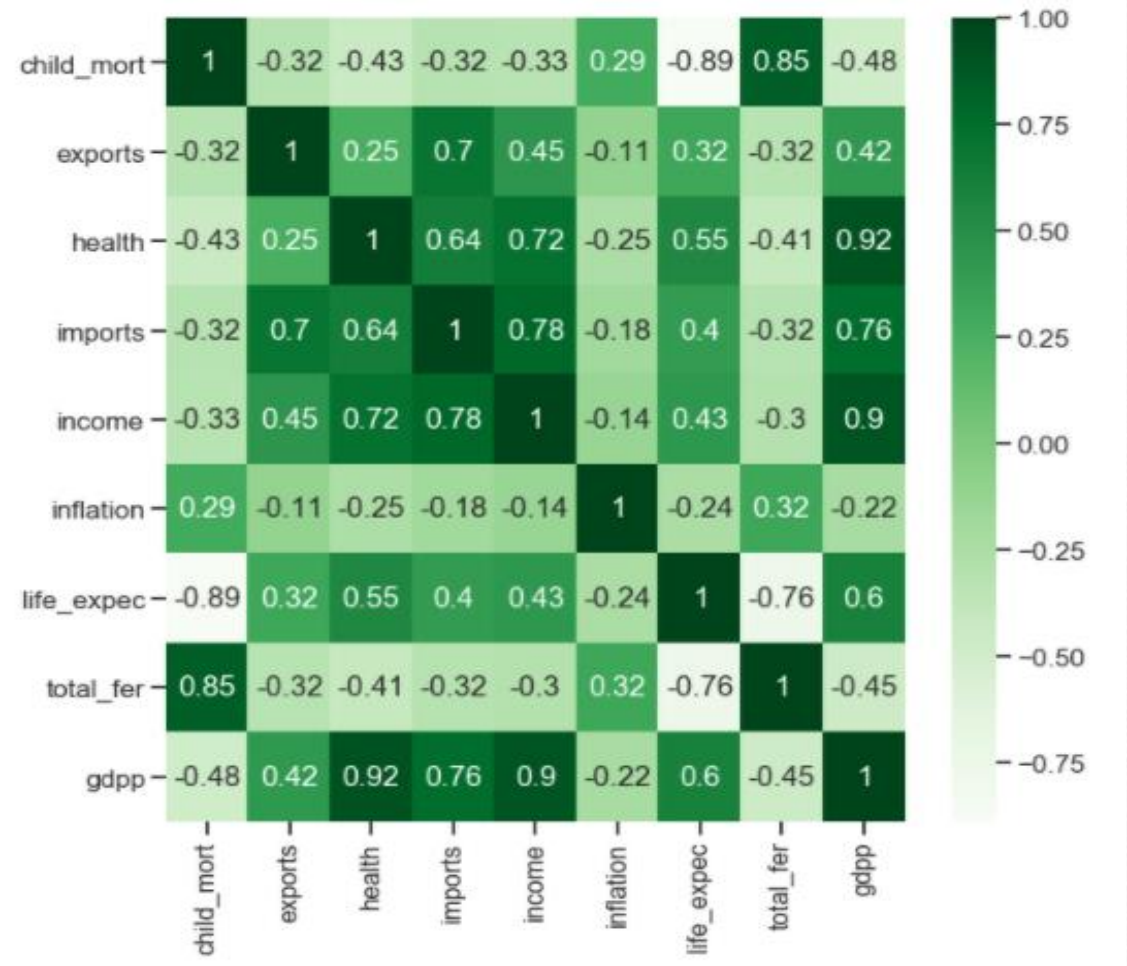
Outliers Analysis

- Plotted a boxplot for all the features to understand the existence of outliers in the dataset.
- The inflation boxplot has less quartiles when compared to others.
- Total_no_of_children_per_women has outliers on the bottom of the boxplot.



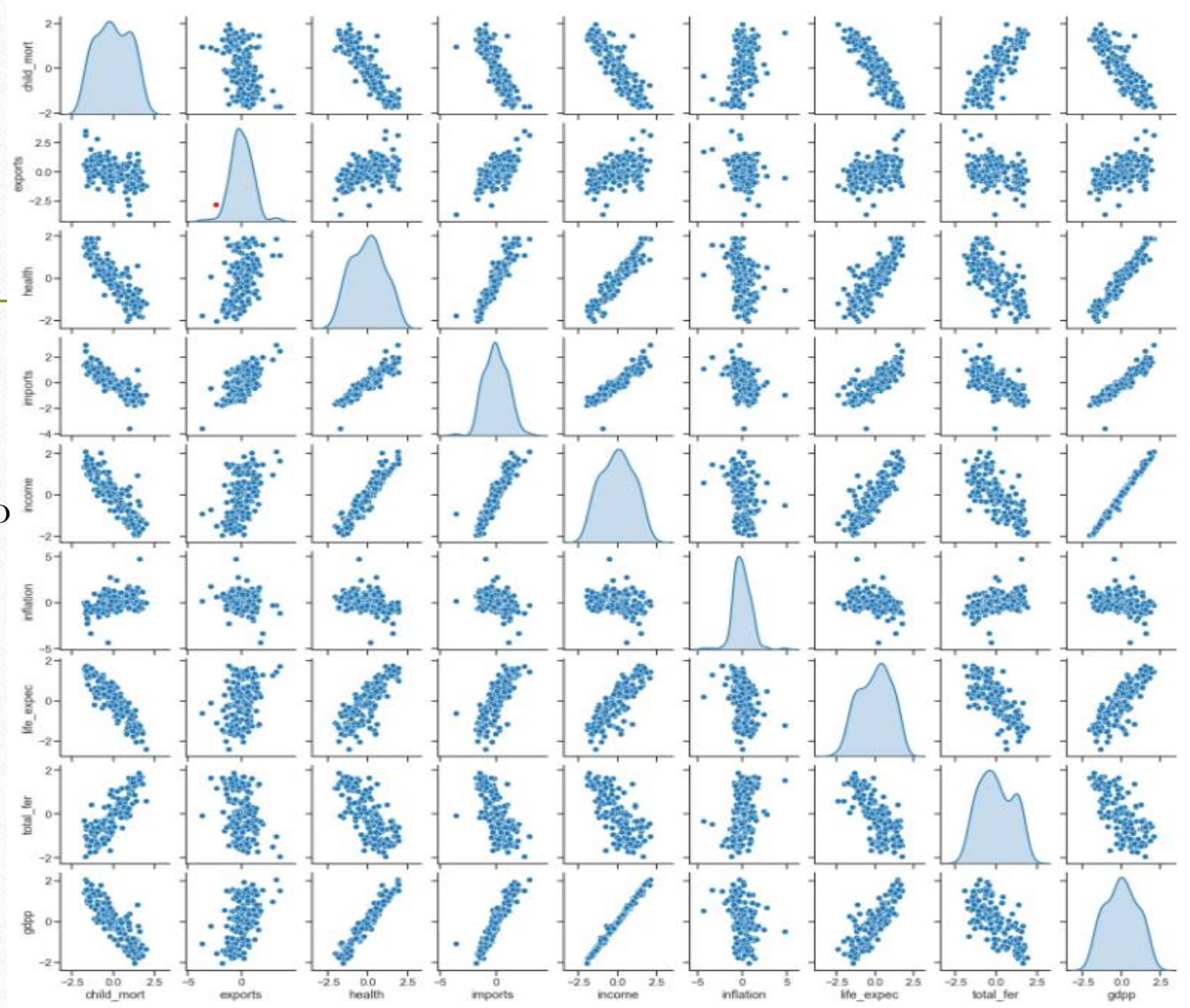
Correlation

- Plotting a heat map to find the correlation between the variables on the dataset.
- As a result some variables are high positive correlation and also high negative correlation.
- Thus we can say the dataset is having multicollinearity.



Pair plot after Rescaling the data

- Plotted a pair plot on the dataset after rescaling .
- Rescaling helps to convert the data into a specific range which helps the accuracy in clustering algorithms.



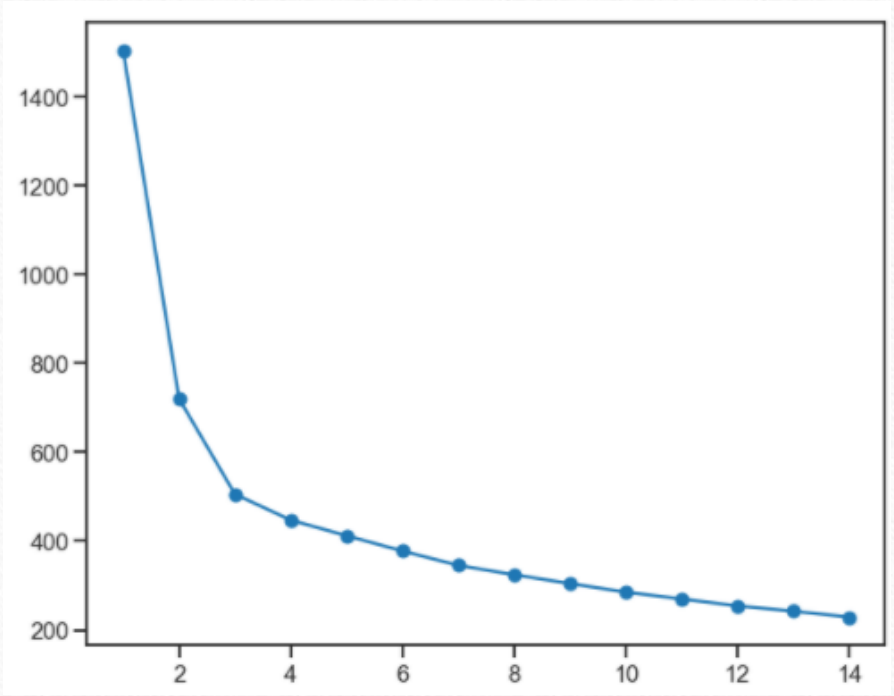
K-Means Clustering

To find Optimal no of clusters we use

- SSD/Elbow curve method.

we have to select the value of k at the "elbow" ie the point after which the distortion/inertia start decreasing in a linear fashion. Thus for the given data, we conclude that the optimal number of clusters for the data is **3**.

Hence k-value = 3.



K-Means Clustering

- **Silhouette Analysis**

  We have found few silhouette scores of the given dataset,

  For n_clusters=2, the silhouette score is 0.4083553693333196

  For n_clusters=3, the silhouette score is 0.34850251239362645

  For n_clusters=4, the silhouette score is 0.27301488301380217

  For n_clusters=5, the silhouette score is 0.24083623893543307

  For n_clusters=6, the silhouette score is 0.26741737502135443

  For n_clusters=7, the silhouette score is 0.2190593898752462

  For n_clusters=8, the silhouette score is 0.25692610741596333

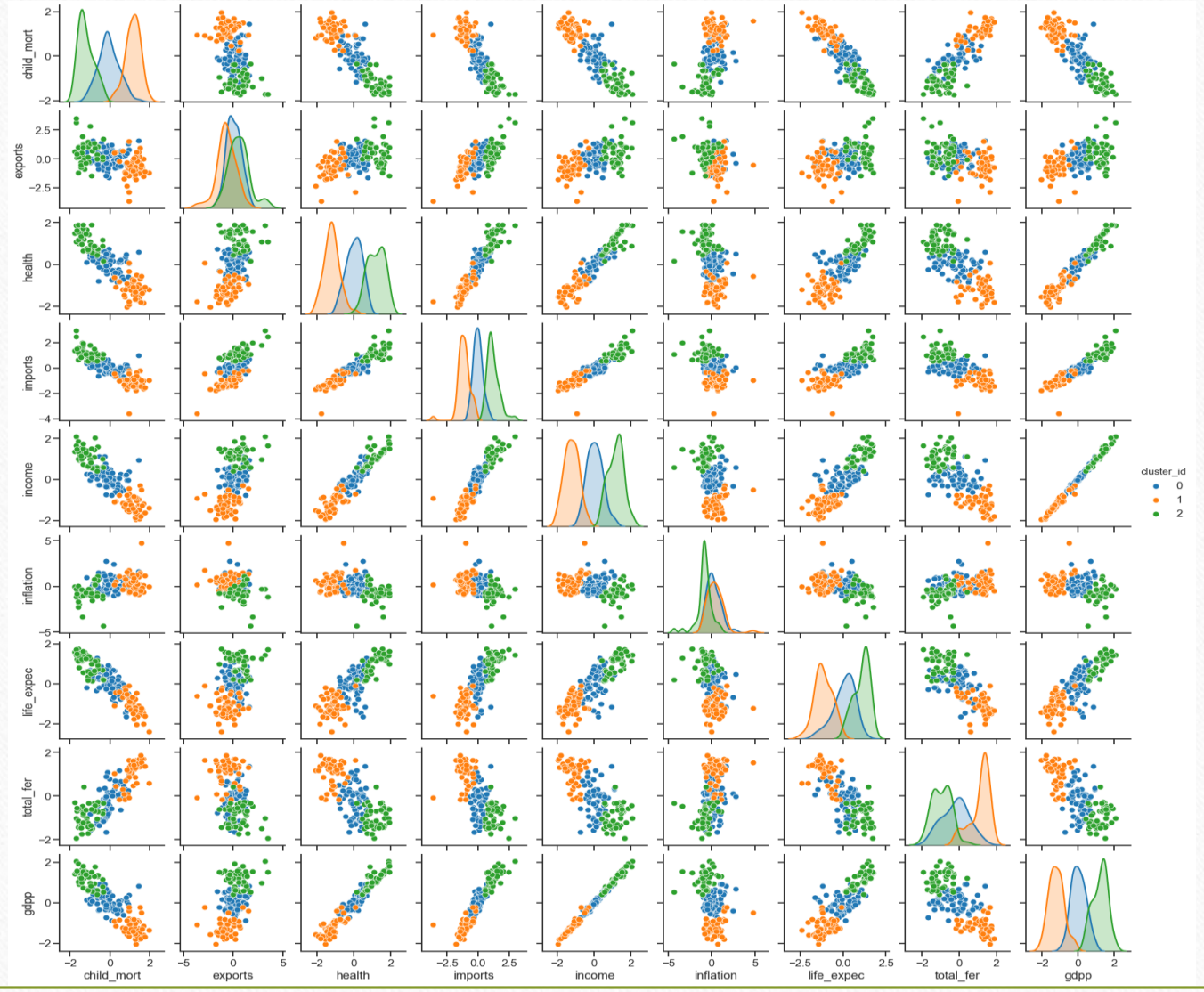
  According to above scores, we can conclude n_cluster = 3 gives a good score when compared to others.

  Hence we can assume K-value = 3

Pair plot with cluster id

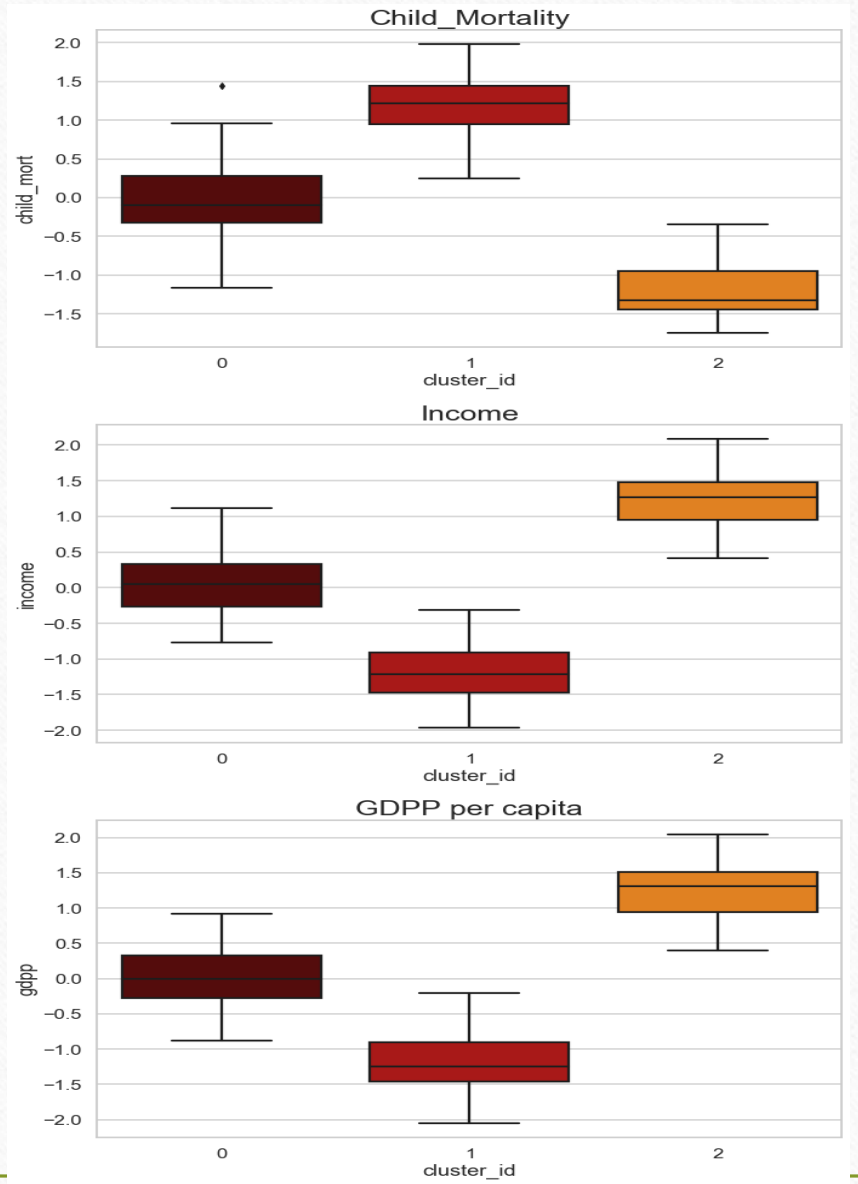
- Plotting a pair plot among all the variables with respect to cluster id.

- Blue – cluster 0
- Orange – cluster 1
- Green – cluster 2



Box plot comparison among Child_Mortality, Income & GDPP (K-Means Clustering)

- Cluster id 0 behaves normal in all features
- Cluster id 1 has High child mortality and also less income & GDPP per capita
- Cluster id 2 has the low child mortality and high income & GDPP per capita.
- So we can consider the list of countries under Cluster id 1
- There are 50 countries in cluster id 1 which needs immediate aid of funding.

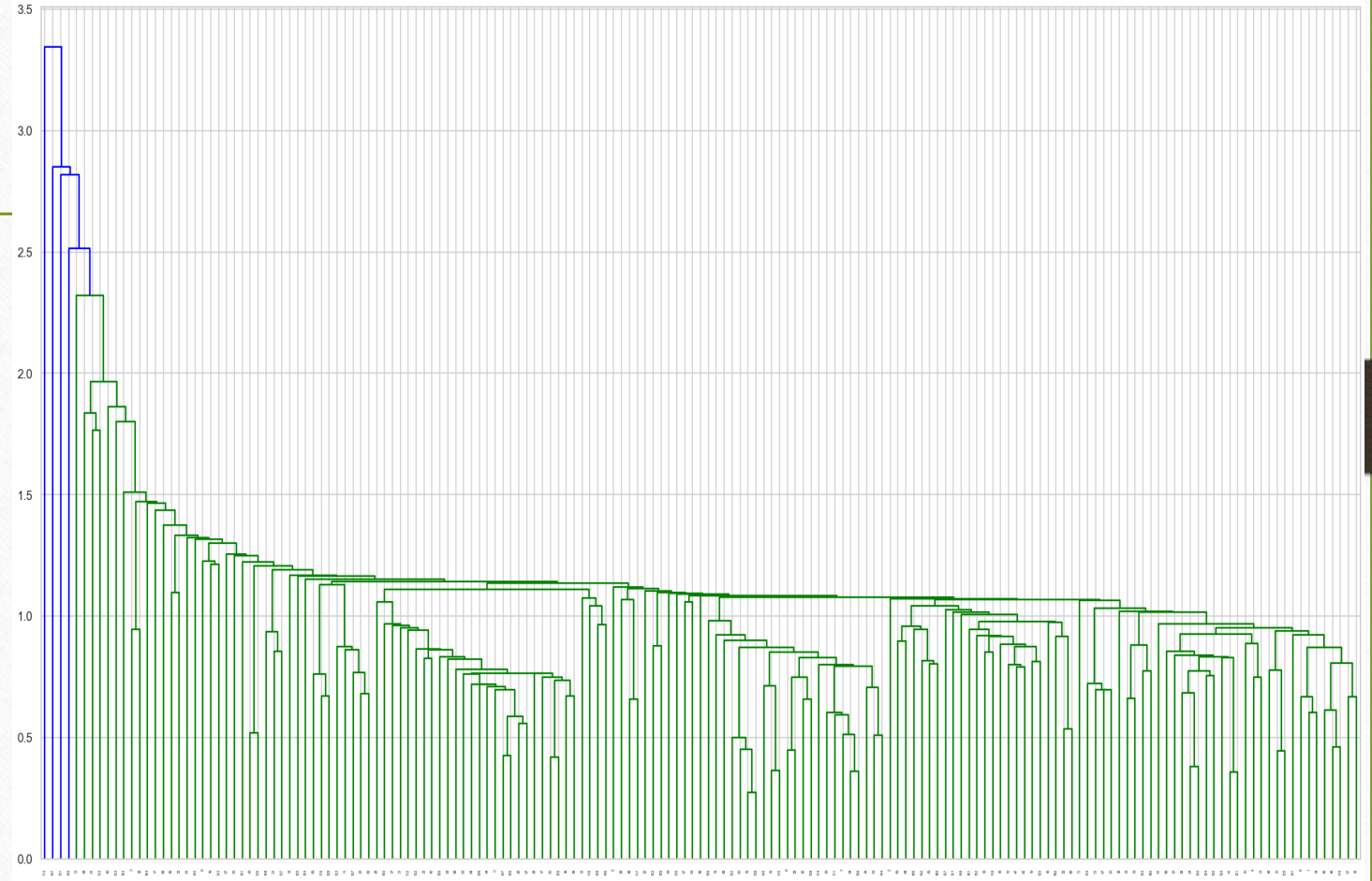


Hierarchical Clustering

- Hierarchical cluster analysis is an algorithm that groups similar objects into groups called clusters.
- In this method we use Single linkage, Complete linkage and plot Dendrogram to find the optimal no of clusters.

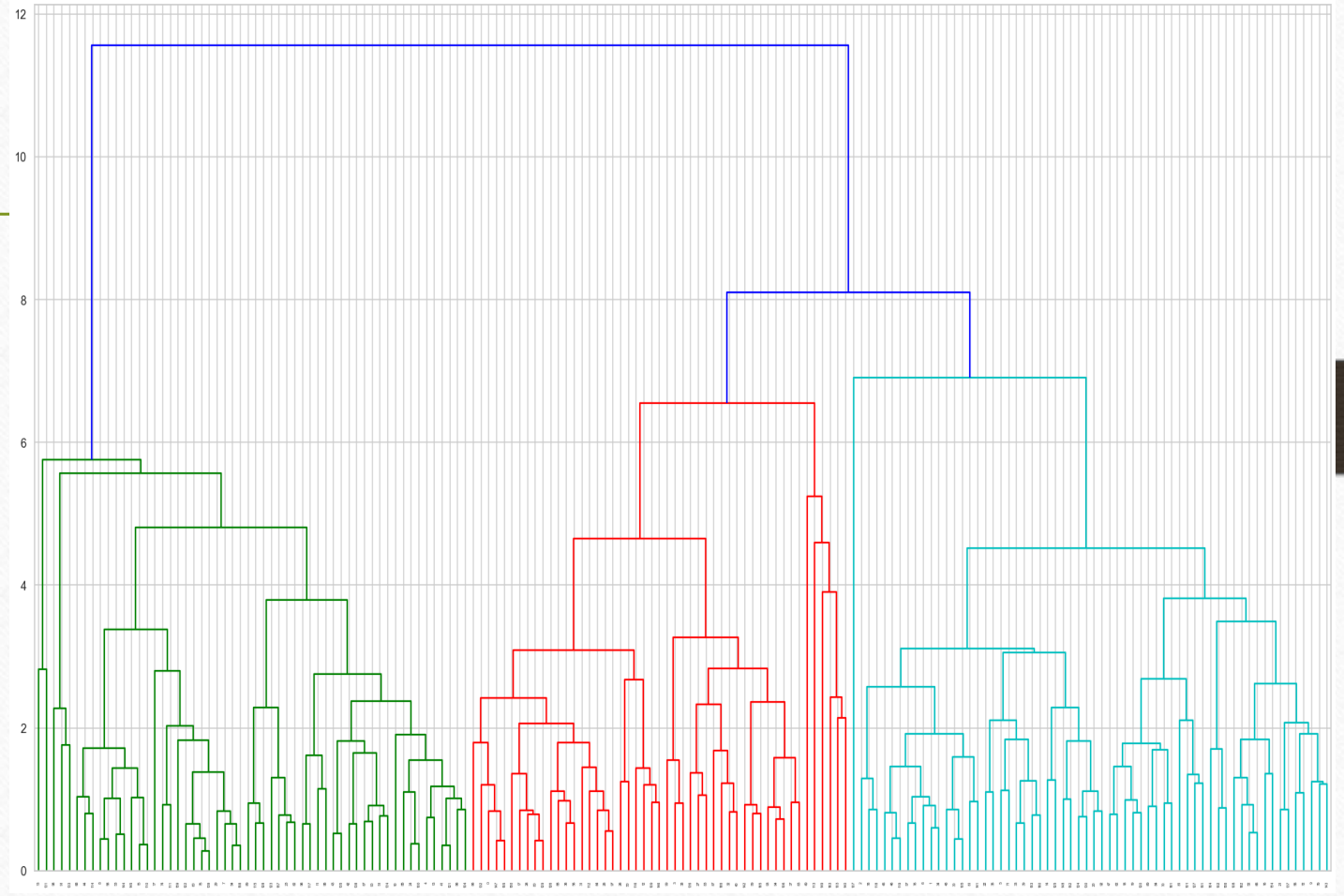
Single Linkage

- In this method we can able to graph the dendrogram using single linkage.
- But unfortunately it's not quite visible and doesn't suits for our dataset.
- Also its difficult to cut the tree in a threshold value. Hence we will use complete linkage method.



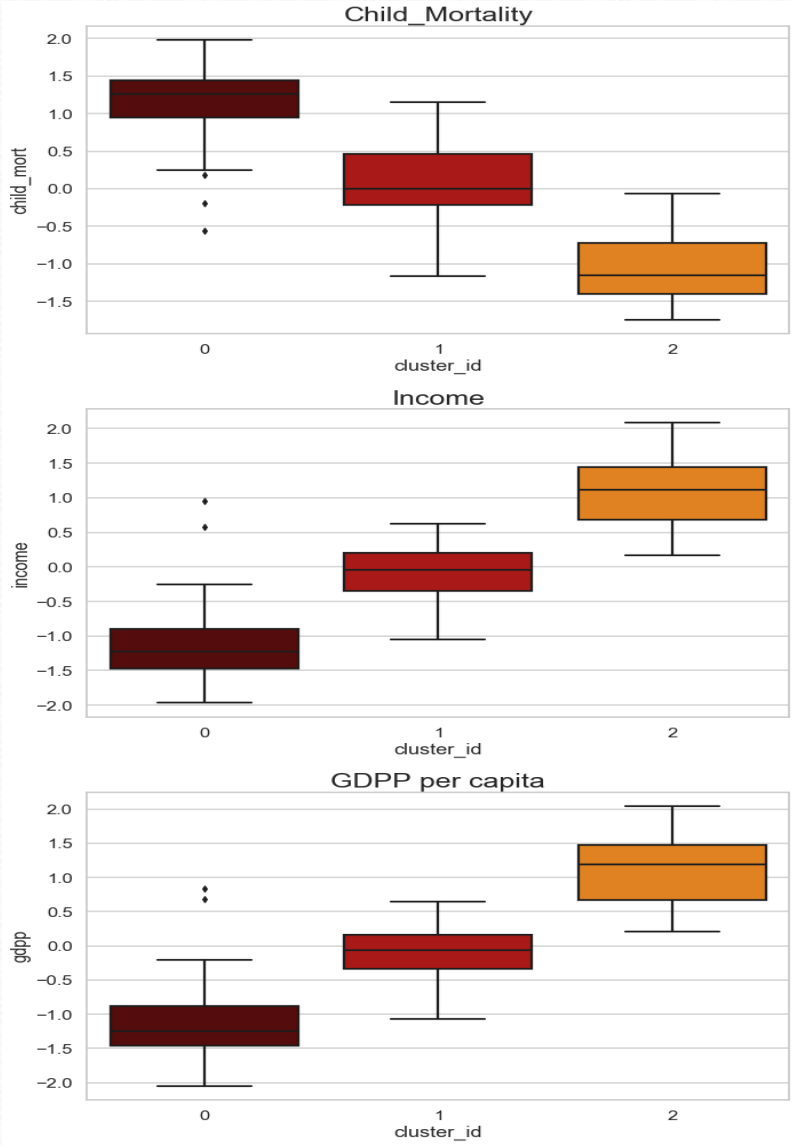
Complete Linkage

- In this graph it is very much visible and helps us to deicide the no of clusters.
- We can able to cut the tree at a threshold value
- We can cut at value 3 which gives us 3 clusters.



Box plot comparison among Child_Mortality, Income & GDPP (Hierarchical Clustering)

- Cluster id 0 has the high child mortality and also less income & GDPP per capita
- Cluster id 1 behaves normal in all features
- Cluster id 2 has the low child mortality and high income & GDPP per capita.
- So we can consider the list of countries under Cluster id 0
- There are 49 countries in cluster id 0 which needs immediate aid of funding.



Insights from both clustering methods.

- By applying both K-Means and Hierarchical Clustering methods on the given dataset. We found 50 countries which need aid (K-Means Clustering) and 49 countries which need aid by performing (Hierarchical Clustering) respectively.
- we have done the experiment with the presence of outliers. If we exclude the outliers we may lose some important information's on the dataset. So we have executed the model with the presence of outliers.
- After analyzing both the methods of clustering, I would choose to go with Hierarchical Clustering which results 49 countries as it gave more accurate numbers when compared to K-Means.

Conclusion

After executing both K-Means and Hierarchical Clustering methods, we have found that the results from Hierarchical Clustering method is more accurate and also this method fulfills the business requirement.

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