Cluster Assignment

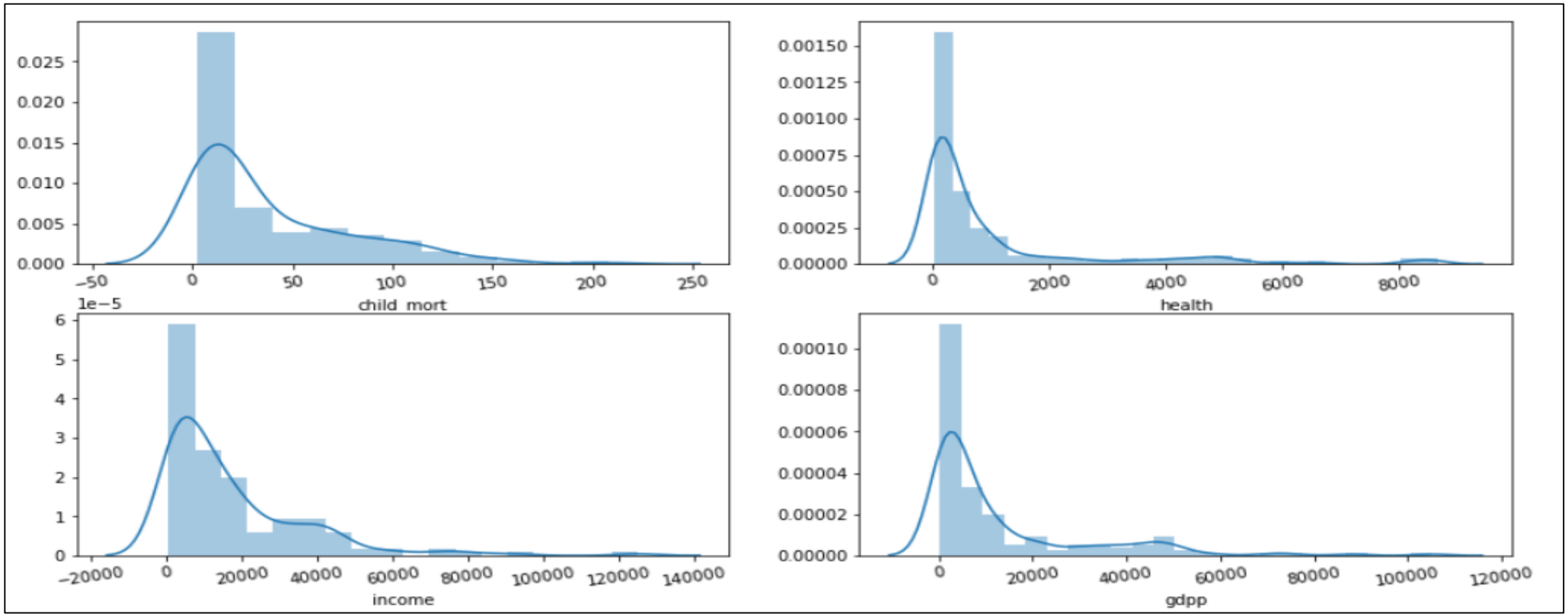
Financial Aid

1. Data Quality Checks and Conversion of Percentage

- First we have taken a detailed analysis of the data set
- There are no NULL values
- There are three columns which are in percentage and I have converted it into value terms as these were the percentage values of total GDPP (Exports, Imports and Health)

2.EDA (Univariate and Bivariate Analysis)

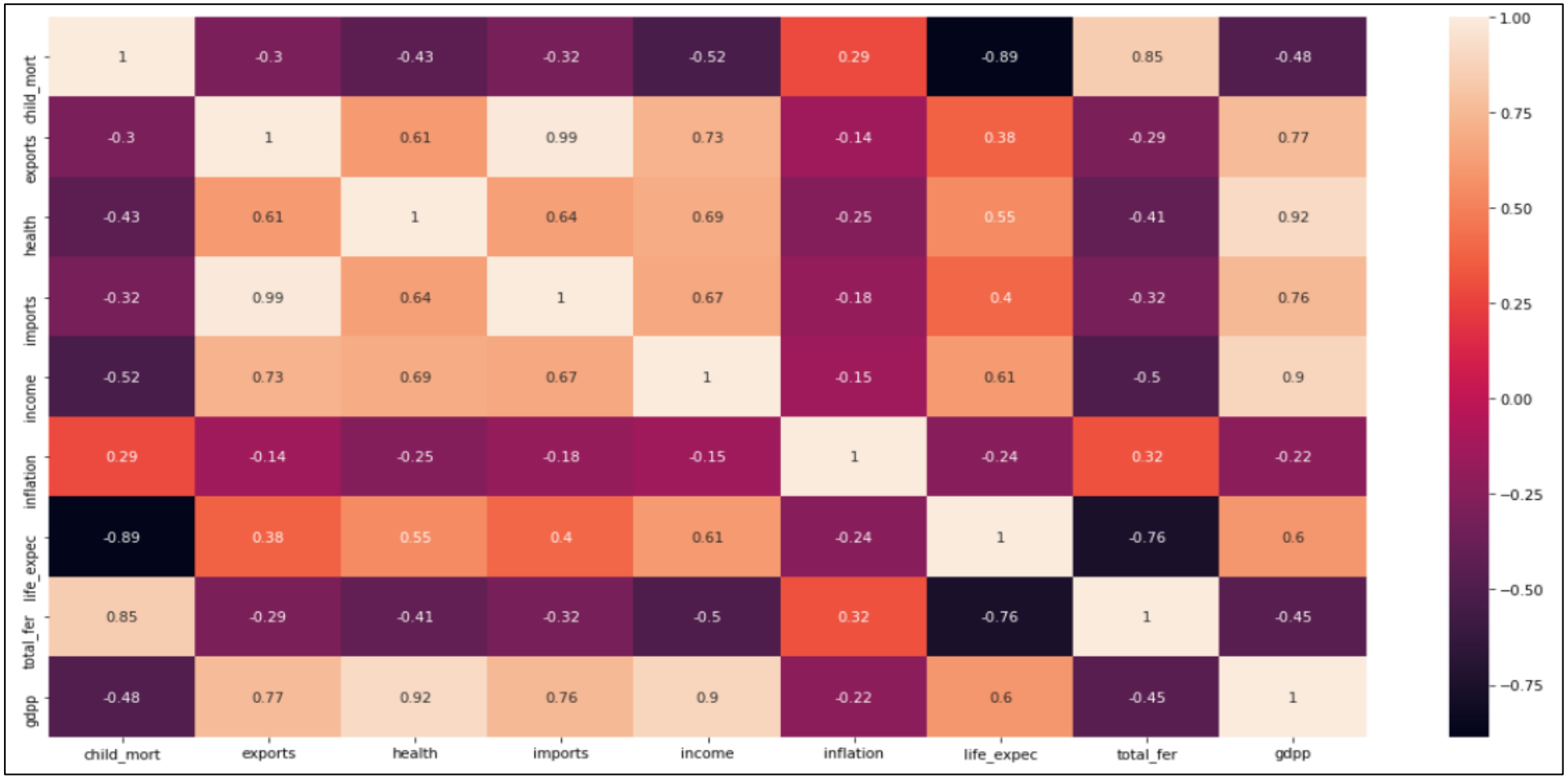
Univariate Analysis



- I have taken the Univariate Analysis on ‘child_mort’, ’health’, ’income’, ‘gdpp’ and plotted a Histogram to check the distribution of error terms
- From the above distplot, we can infer that all the variables are showing similar types of variations and they are uniformly distributed

2.EDA (Univariate and Bivariate Analysis)- Continued

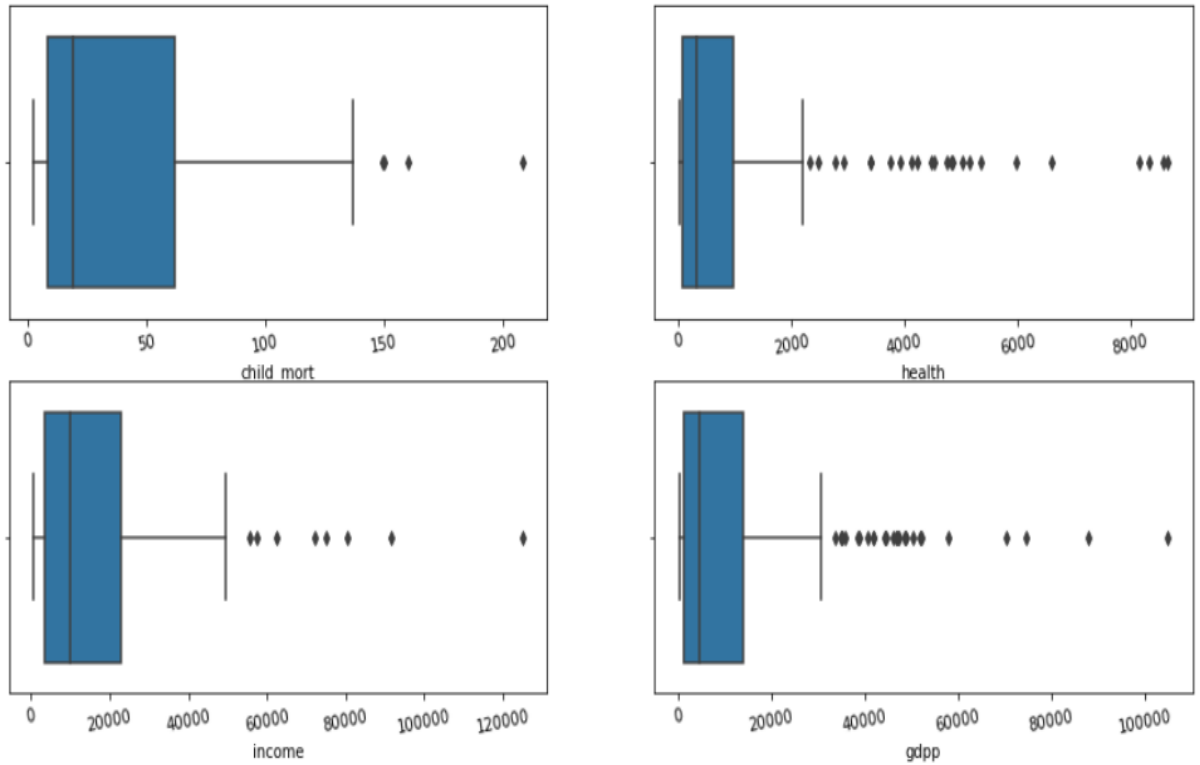
Bivariate Analysis



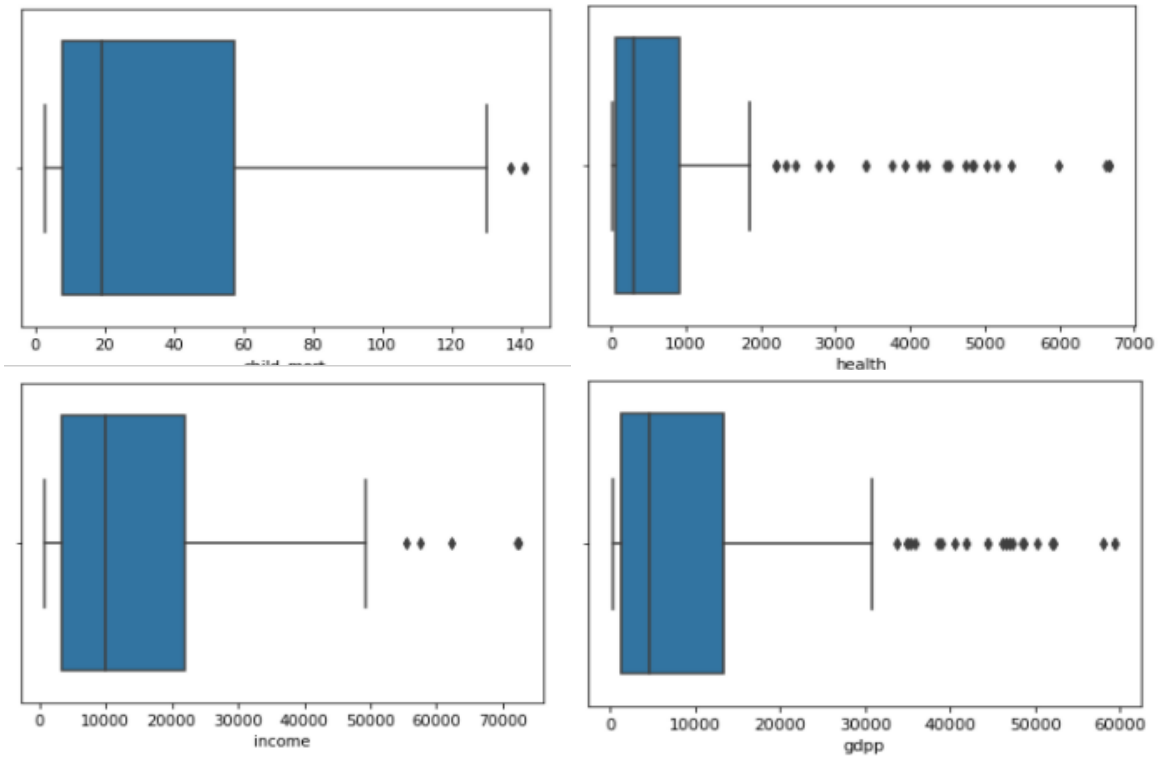
- I have taken the Bivariate Analysis on ‘child_mort’, ’health’, ’income’, ‘gdpp’ ,’exports’,’import’,’inflation’,’life_expec’, ‘total_fer’and plotted a Heat map to check the distribution of error terms
- The lighter shade corelates to higher correlation. For e.g Higher the 'GDPP', higher is the 'health', higher the 'GDPP', higher is the 'income'
- The darker shade correlates to negative correlation. That means if value of one variable increases, the value of other variable will decrease and vice versa. For e.g 'Child mort' and 'life_expec', 'total_fer' and 'child_mort'

3.Outlier Analysis

Outlier Analysis on ‘child_mort’, ’health’, ’income’, ‘gdpp’



Before Capping



After Capping

- Here I can see that there are number of outliers in 'health' , 'income' , 'gdpp' and few outliers in 'child_mort’
- I have done capping through soft range to replace extreme values with acceptable limits
- Also I have avoided hard range so that I should not drop any countries which might need actual aid

4.Scaling the Data set

Standardised Scaling

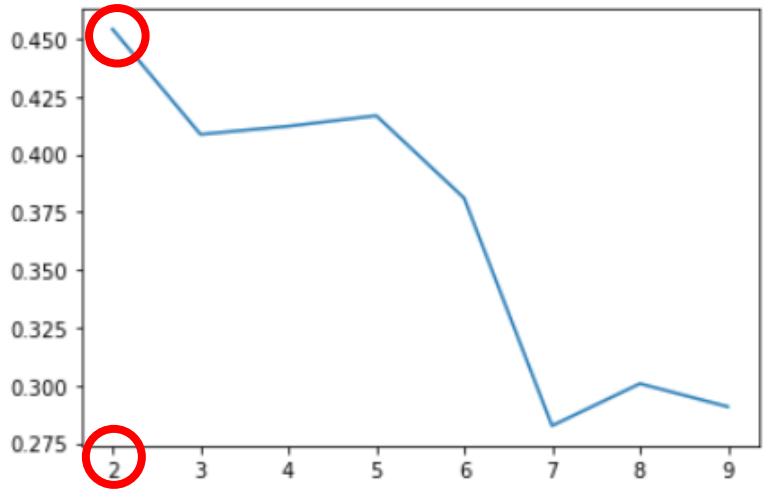
- I have done Standardised scaling on the data set as there are extreme outliers
- I have done rescaling as without rescaling the regression model might be very large or small as compared to other co-efficient and will lead to confusion during model evaluation

5.Hopkins, Silhouette and Elbow Curve

Hopkins Statistics

- By doing Hopkins Statistics, I can find that the value is close to 1, so the data is highly clustered as per the rules of Hopkins statistics

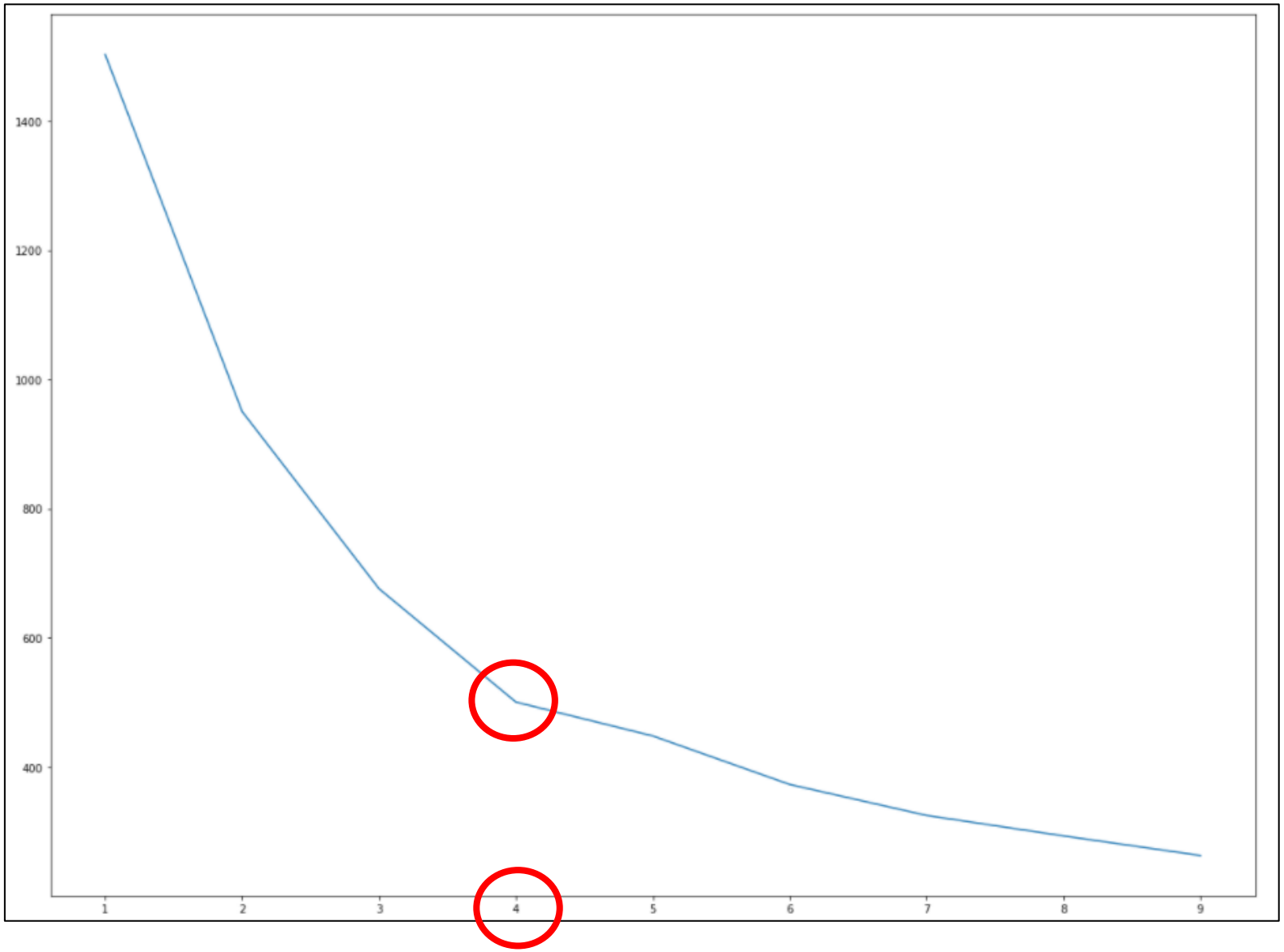
Silhoutte Plot



- The Silhouette plot shows that the silhouette coefficient is highest when k=2 suggesting the optimal number of clusters

5.Hopkins, Silhouette and Elbow Curve (Continued)

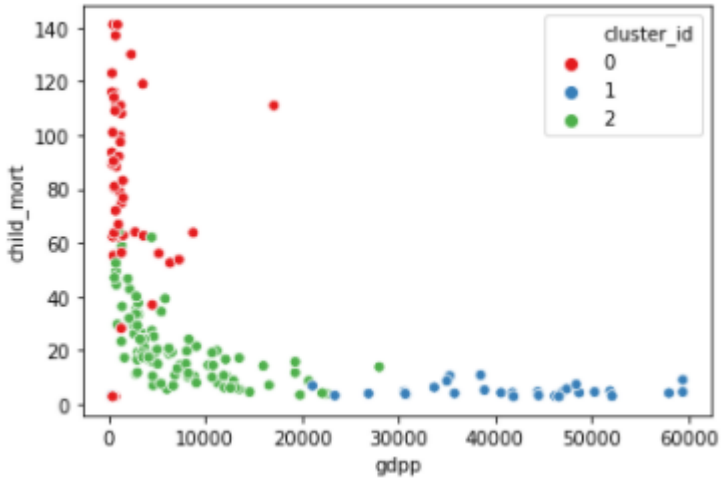
Elbow Curve



- Here I have to select the value of k at the 'elbow' that is the point after which the distortion/inertia start decreasing in a linear fashion.
- Here I can conclude that the optimal number of clusters for the data is 4

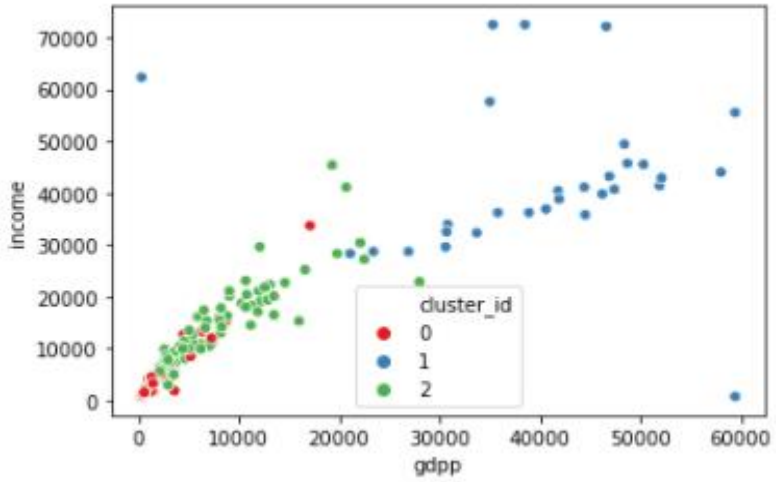
6.K-Means Application and Visualisation

K-Means Visualisation



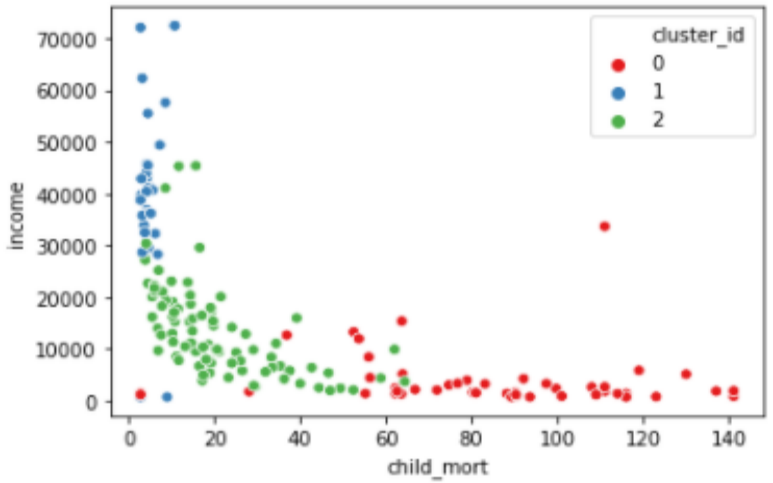
When drawn scatterplot on ‘child_mort’,’gdpp’, I have the following observations :-

a) For cluster 2, as ‘gdpp’ increases, there is not much increase in ‘child_mort’ however when the ‘gdpp’ is below 5000, ‘child_mort has gone upto more than 60 also
b) For cluster 0,when ‘gdpp’ is low (below 5000), the ‘child_mort’ is very high(has gone upto more than 140) indicating a very high mortality rate in children with the exception of 1 outlier with high ‘gdpp’ also
c) For Cluster 1, here the ‘gdpp’ of the countries are the highest among the 3 clusters and the ‘child_mort’ is also very low(close to 20) showing low mortality in children



When drawn scatterplot on ‘income’,’gdpp’, I have the following observations :-

a) For cluster 2, as ‘gdpp’ increases, there is an increase in ‘income’. Also I can see some outliers as ‘gdpp’ touches 20000
b) For cluster 0, when ‘gdpp’ is low (below 5000), the ‘income’ is low. Also there is one outliers as ‘gdpp’ increases towards 20000
c) For Cluster 1, here the ‘gdpp’ of the countries are above 20000 and the corresponding income is also high starting from 30000 which concludes that higher the ‘gdpp’, higher is the ‘income’. Also here there are number of outliers with extremely high ‘income’

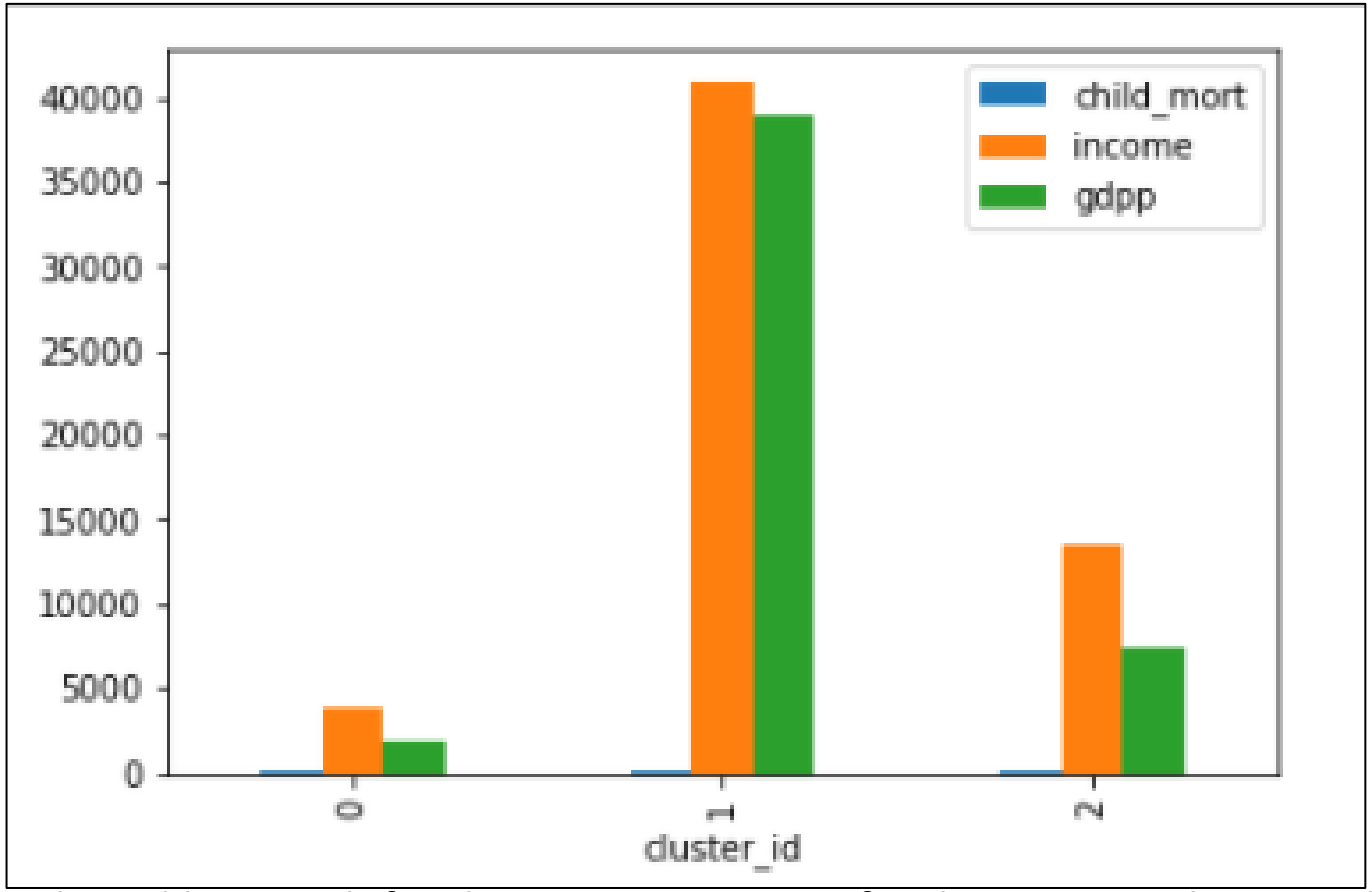


When drawn scatterplot on ‘income’,’child_mort’, I have the following observations :-

a) For cluster 2, as ‘child_mort’ increases to 60, the ‘income’ is quite low indicating that income is inversely proportional to ‘child_mort’
b) For cluster 0,mortality rate increases with low ‘income’. Here there is an 1 outlier between 30000 and 40000 ‘income’
c) For Cluster 1, the mortality rate is the lowest in ‘child_mort’ as the ‘income’ is quite high as compared to the other two clusters

6.K-Means Application and Visualisation (Continued)

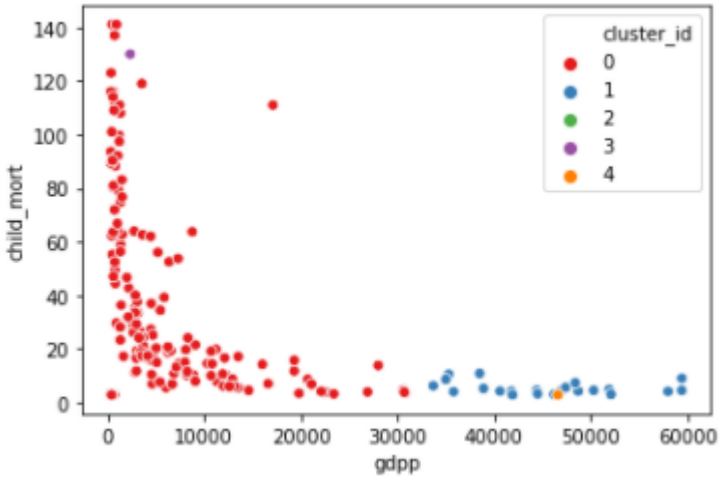
K-Means Application



- I have plotted bar graph for the K-means status of 3 clusters namely 0,1 and 2
- Cluster 0 is having the lowest ‘income’, low ’gdpp’ and high ‘child_mort’ (we can see this from the mean also)
- By analysing the clusters on the basis of K-Means, we can conclude that countries such as 'Burundi','Liberia' , 'Congo,Dem. Rep' , 'Niger' & 'Sierra Leone' are in dire need of AID

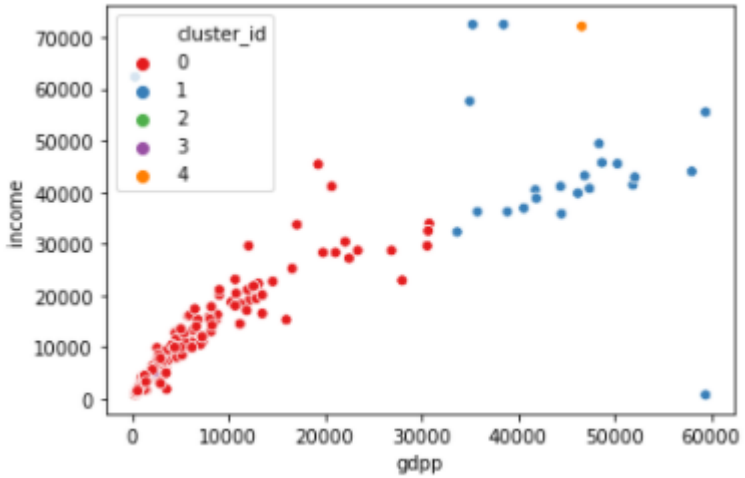
7.Hierarchical Clustering

Hierarchical Clustering Visualisation



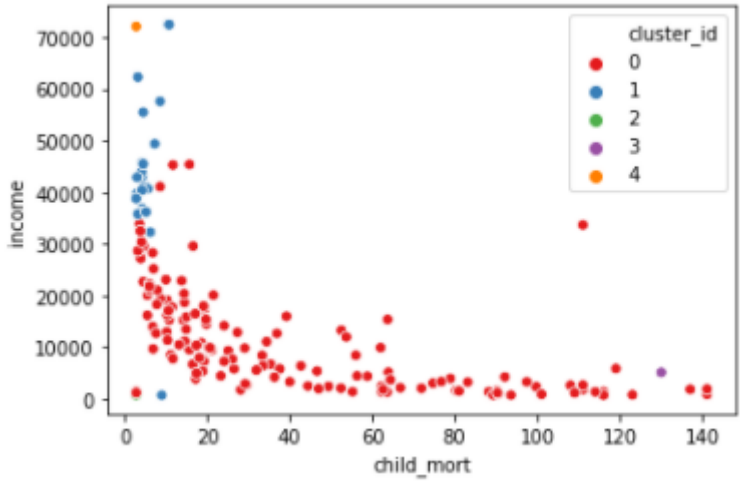
When drawn scatterplot on ‘child_mort’,’gdpp’, I have the following observations :-

a) Here only cluster 0 and cluster 1 are having maximum data points
b) In cluster 0, ‘child_mort’ is higher when ‘gdpp’ is low. Also there is an outlier when ‘gdpp’ is close to 20000
c) In cluster 1, ‘child_mort’ is minimum as ‘gdpp’ is quite high among all the clusters
d) For clusters 2 and 4, there is just 1 data point each having low ‘child_mort’
e) For cluster 3 also, there is one data point having high ‘child_mort’



When drawn scatterplot on ‘income’,’gdpp’, I have the following observations :-

a) Here only cluster 0 and cluster 1 are having maximum data points
b) In cluster 0, ‘income’ is higher with the increase in ‘gdpp’. There are some outliers when ‘gdpp’ is above 20000
c) In cluster 1, ‘income’ is quite high as ‘gdpp’ is also quite high among all the clusters.There are some outliers when ‘gdpp’ crosses 30000
d) For clusters 2,3 and 4, there is just 1 data point each however that single datapoint in cluster 4 is having high ‘income’

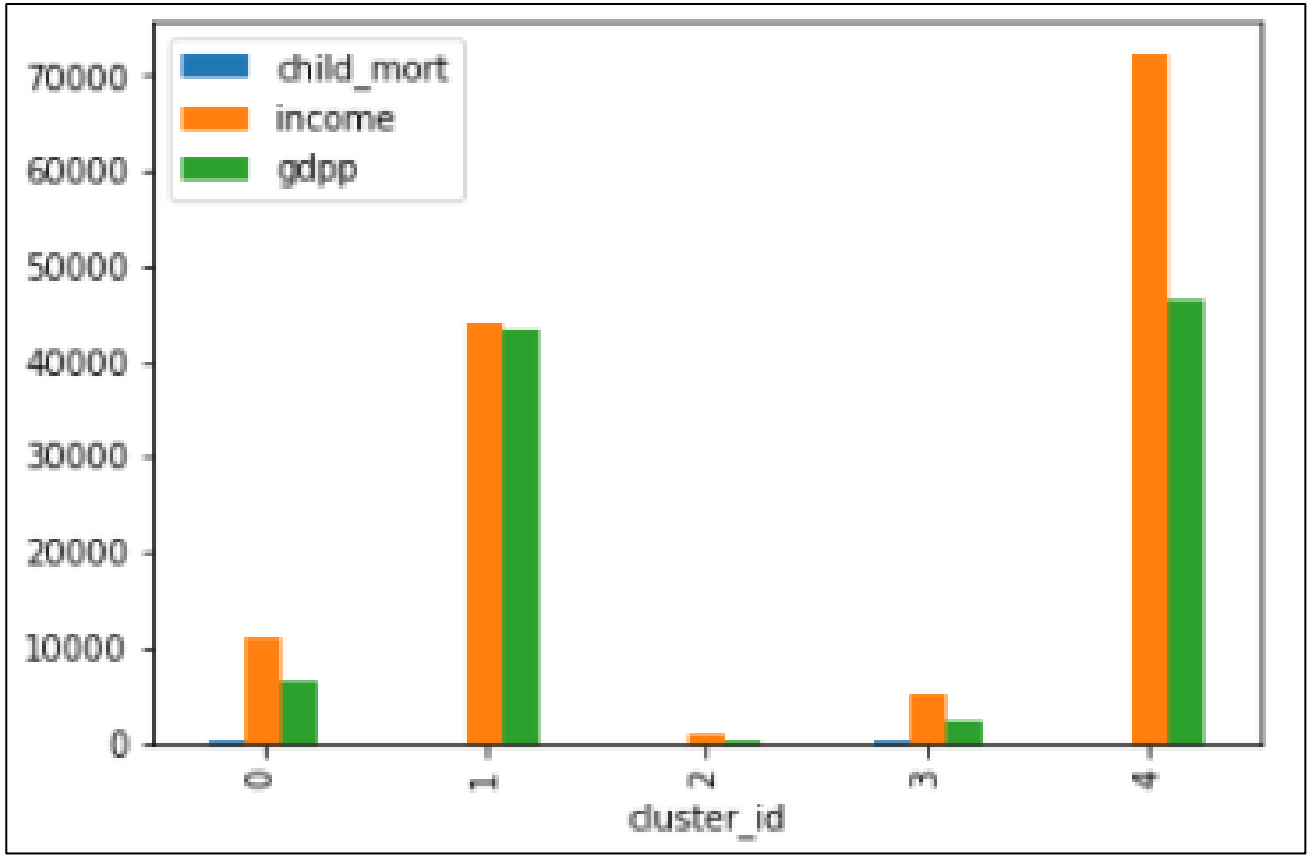


When drawn scatterplot on ‘income’,’child_mort’, I have the following observations :-

a) Here only cluster 0 and cluster 1 are having maximum data points
b) In cluster 0, ‘child_mort’ is higher when ‘income’ is low. Also there is an outlier when ‘child_mort’ is more than 150
c) In cluster 1, ‘child_mort’ is minimum as ‘income’ is quite high among all the clusters
d) For clusters 2,3 and 4, there is just 1 data point however that single datapoint in cluster 4 is having high ‘income’

7.Hierarchical Application (Continued)

Hierarchical Application



- I have plotted bar graph for the Hierarchical status of 5 clusters namely 0,1,2,3 and 4
- Cluster 2 is having the lowest ‘income’, low ’gdpp’ and high ‘child_mort’ (we can see this from the mean also)
- By analysing the clusters on the basis of Hierarchical clustering, we can conclude that there is only one country in cluster 2 by the name ‘Luxembourg’ which is in dire need of AID

8.Final Conclusion

Final list of Countries

Since we have a precise list of countries from K-Means clustering , so I have considered the following list of atleast 5 countries which are in dire need of AID

- 'Burundi'
- 'Liberia'
- 'Congo Dem. Rep'
- 'Niger'
- 'Sierra Leone'