Clustering Assignment

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.

After the recent funding programmes, 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.

We need to categorise the countries using some socio-economic and health factors that determine the overall development of the country. Then suggest the countries which the CEO needs to focus on the most.

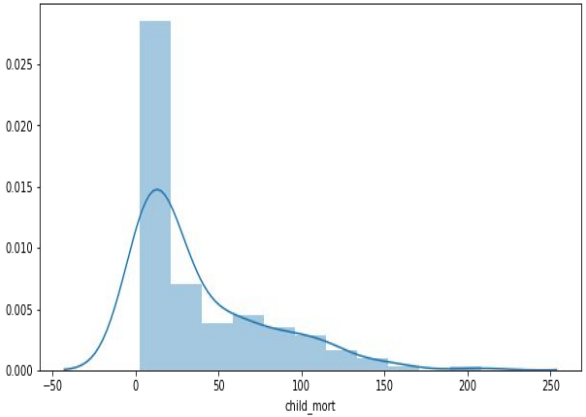
Analysis Approach

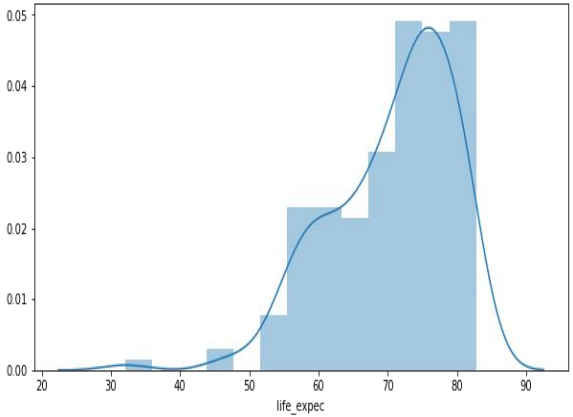
1. Data Quality Check
2. EDA: Univariate / Bivariate Analysis
3. Outlier Treatment
4. Check the cluster tendency: Hopkin's Test
5. Perform Cluster profiling: GDPP, CHILD_MORT, and INCOME
6. Scaling
7. Find the best value of **K**: SSD, Silhouette Score
8. Using the final value of **K**, perform final K Means Analysis
9. Visualize the cluster using a scatter plots
10. Hierarchical Clustering (Single Linkage: Dendrogram and Complete Linkage: Dendrogram)

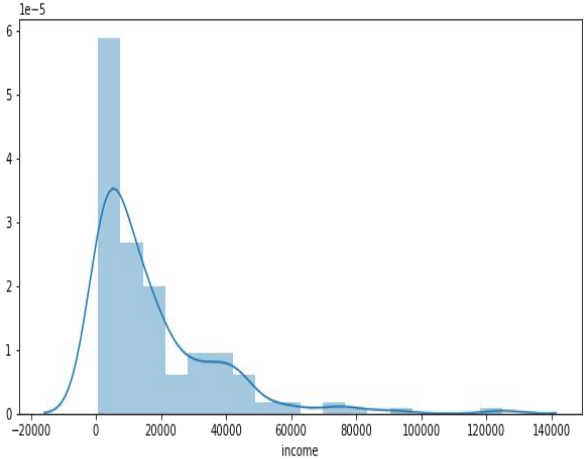
Univariate Analysis

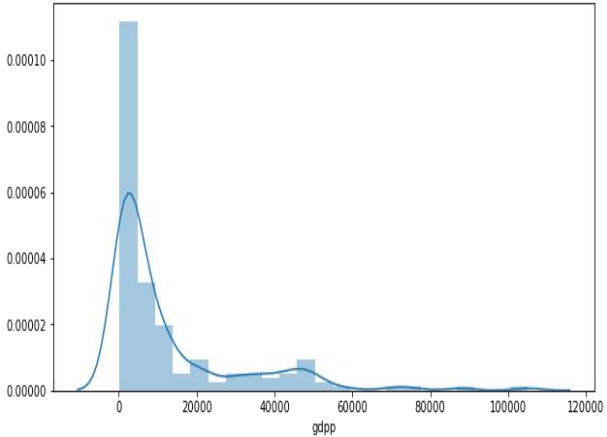
After doing Data Quality check, we found that columns such as exports, health, imports were representing percentage and was converted to respective value form other than that there wasn't any missing values and all the columns did not contain any invalid values

Univariate analysis was done to find the distribution of values through various columns. We can find some distribution plots in the right side of the slide





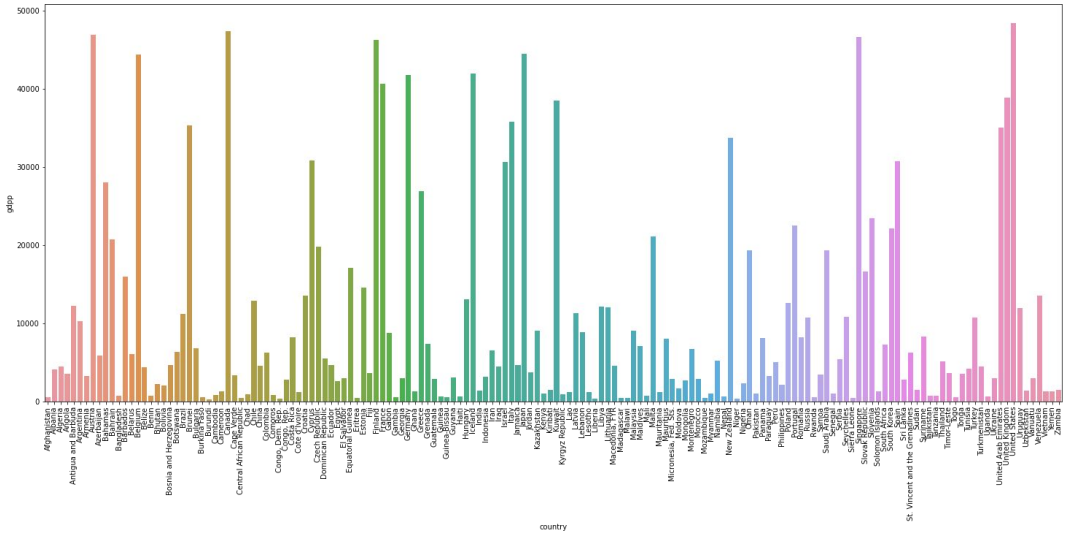


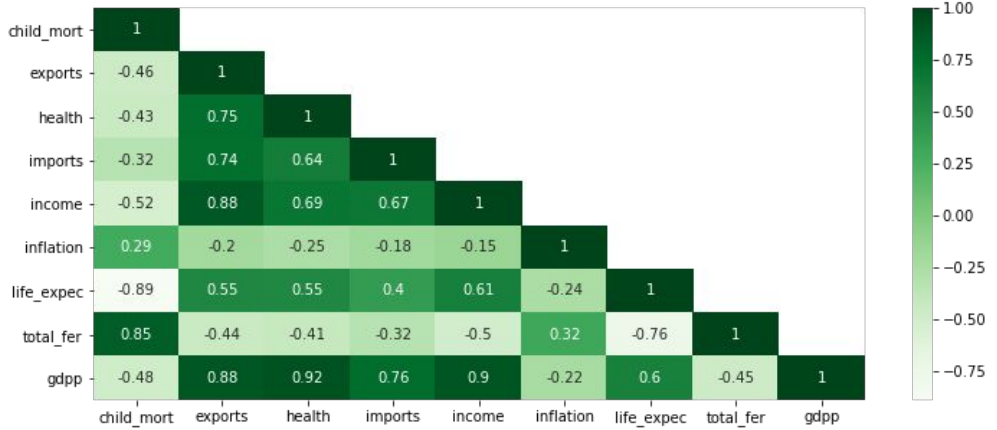


Bivariate Analysis

Bivariate Analysis for numeric columns was done by plotting pair plot and correlation matrix, Here we came to know that ‘gdpp’ column had high correlation with exports, health, imports, income columns

Categorical analysis was done by plotting a bar plot against country column and also we came to know that the countries which had high value of gdpp had low values of child_mort, inflation and total_fer





Outlier Treatment and Cluster Tendency

As the problem statement defines that NGO wants to aid the countries which are direst in need, we can conclude that countries having high gdpp and income values are far better than the countries having low gdpp and income values and also countries having low child_mort are in direst need.

Since the clustering process is sensitive to outliers we removed the top 5 percent values of gdpp column, thus considering bottom 95 percent values.

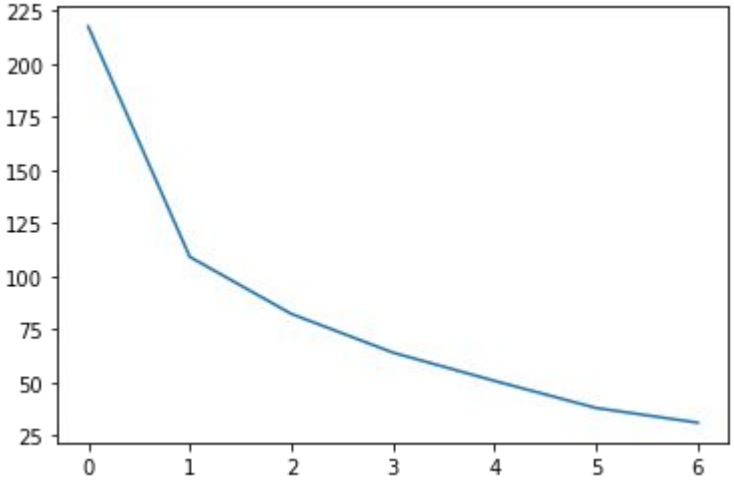
We conducted Hopkins test to calculate the percentage of cluster tendency and got desired output of 96 percent which was satisfied

Best Value of K

To choose the number of clusters we determined K from two methods

1. Elbow Curve
2. Silhouette Score

From the above analysis through plots and scores we have decided to choose the value of K=3.



For n_clusters=2, the silhouette score is 0.5779643249650761
For n_clusters=3, the silhouette score is 0.5317028614846651
For n_clusters=4, the silhouette score is 0.4498424229841292
For n_clusters=5, the silhouette score is 0.4352209089586231
For n_clusters=6, the silhouette score is 0.44360500908333955
For n_clusters=7, the silhouette score is 0.46311401915249123
For n_clusters=8, the silhouette score is 0.46611293957585964

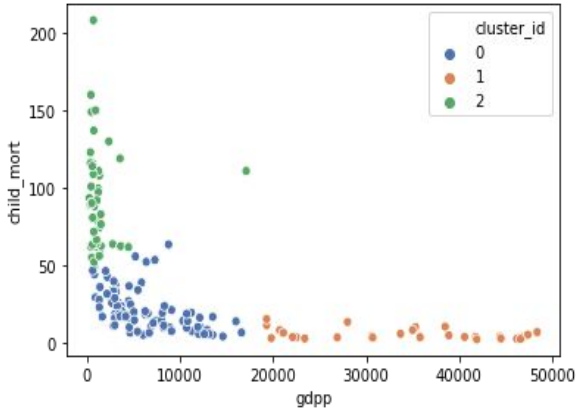
K Means Analysis

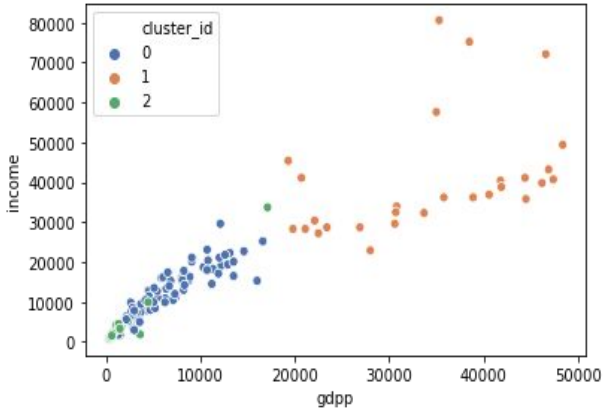
Here for modelling we have chosen 3 columns as ‘gdpp’, ‘child_mort’, ‘income’

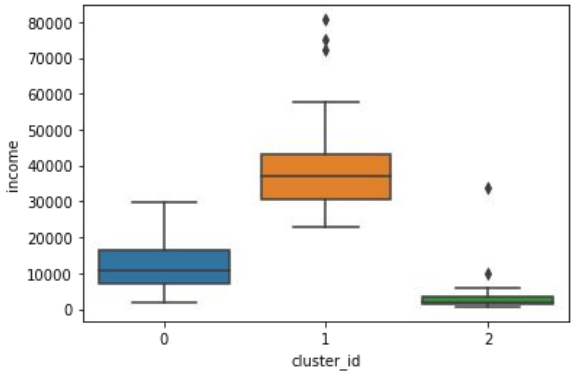
From the above value of K=3 we have chosen 3 clusters for modelling.

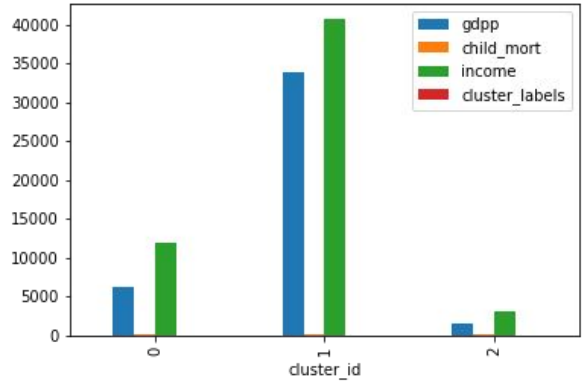
We found that cluster_id = 2 contains the countries which are in dire needed for aid

For cluster_id 0 and 1 contains the countries which have average and high gdpp and income respectively









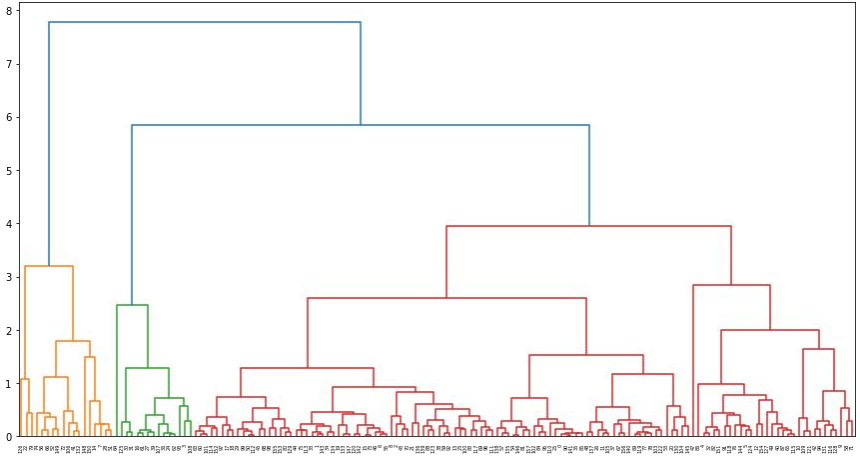
Hierarchical Clustering

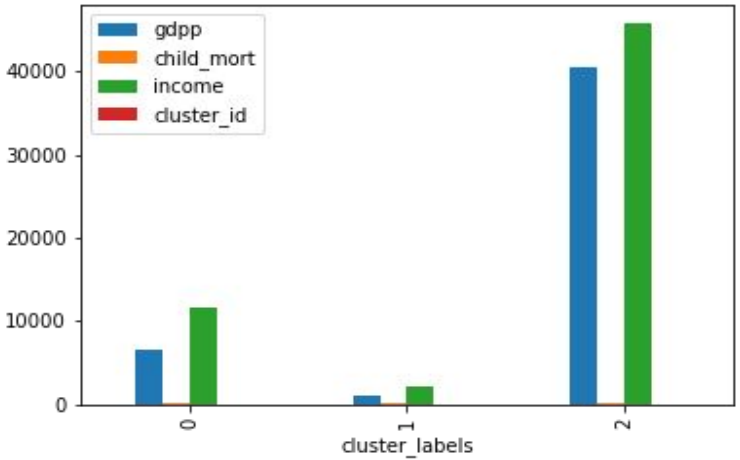
We have use two linkage for interpreting the dendrogram

1. Single Linkage
2. Complete Linkage

Since Complete Linkage had clarity we chose the value of K=3 by interpreting the dendrogram

We found that cluster_labels = 1 contains the countries which are in dire needed for aid





Final Results

From the Above two models we chose the results of the K Means clustering and reporting the 5 countries which are in direst need of aid

1. Burundi
2. Liberia
3. Congo, Dem. Rep.
4. Niger
5. Sierra Leone