International Institute of Information Technology Bangalore

upGrad

HELP International NGO

Clustering Assignment

By : Vyas Bhaumik Hemantkumar

International Institute of Information Technology Bangalore

upGrad

Business Problem

- **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, 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.**

International Institute of Information Technology Bangalore

upGrad

Business Objective

- **To categorize the countries for Financial Aid by Help International NGO on the basis of some socio-economic and health factors that determine the overall development of the country.**

International Institute of Information Technology Bangalore

upGrad

Methodology

- Nine features as socio-economic factors are provided for each of the 167 countries. These features are collinear.
- Clean the data and perform EDA. We will use derived metrics where suitable.

Example : % Health converted to Health per person.

- Data is standardized as features are different units and scale.
- We will attempt to reduce the dimension while retaining the information/ variance.
- From PC converted data, we will attempt to cluster using unsupervised learning techniques like K-means and Hierarchical clustering. Cluster countries based on their socio-economic factors.
- We will treat the outliers i.e. countries with very high or very low development characteristics to enable clustering algorithm to work.
- Once under-developed country cluster is identified we will use is centroid/ mean/ characteristics to find the most under developing countries which require aid the most.
- A comparison of K-means and hierarchical clustering will be done and if variations seen will try to explain them.

International Institute of Information Technology Bangalore

upGrad

Business Problem Solving Approach

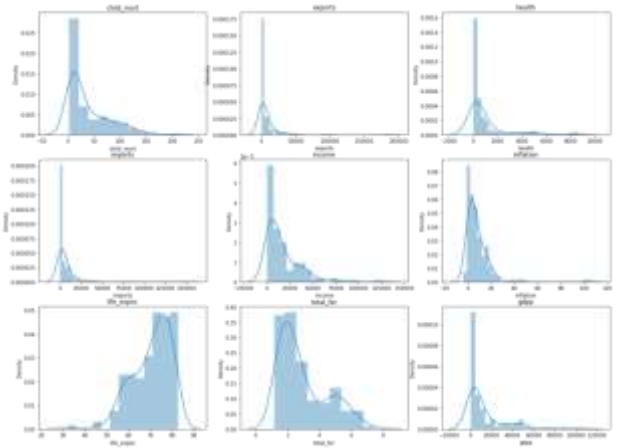
- ❑ Check for missing value and treatment.
- ❑ Check for outlier and treatment.
- ❑ Perform the basic EDA to find the variability and distribution of the data, so as to identify if we need to scale the data.
- ❑ Data Scaling if necessary.
- ❑ Use Hopkins Method to check if the dataset is good enough for a cluster analysis.
- ❑ Using Hierarchical clustering to identify the optimal cluster value.
- ❑ Use Silhouette and Elbow method to validate the optimal cluster values.
- ❑ Use K-Means Cluster method to build the final cluster model.
- ❑ Analyze the cluster that is representing the countries that will solve the Business Problem.
- ❑ Present the final report.

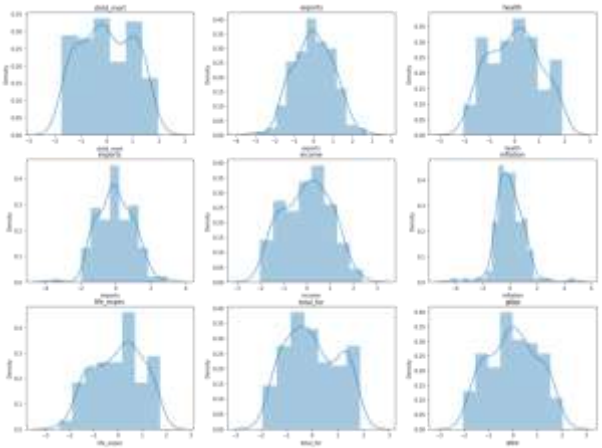
International Institute of Information Technology Bangalore

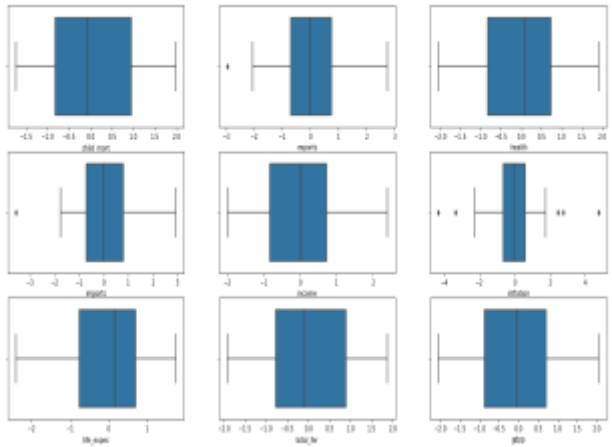
upGrad

Missing Values & Outliers Treatment

- ❑ **Data frame has data about various countries and their socio-economic factors. Few are in % and others in absolute values.**
- ❑ **Data frame has 10 Columns and 167 Rows.**
- ❑ **One variable is 'Object' Type, and rest all are 'Int' or 'Float' type.**
- ❑ **Descriptive Statistics tells us that there is variablity in the data, and will require scaling before model building.**
- ❑ **Plotting all the features to visualize and look their distributions.**
- ❑ **'child_mort', 'exports', 'health', 'imports', 'income', 'inflation', 'life_expec', 'total_fer', 'gdpp'.**





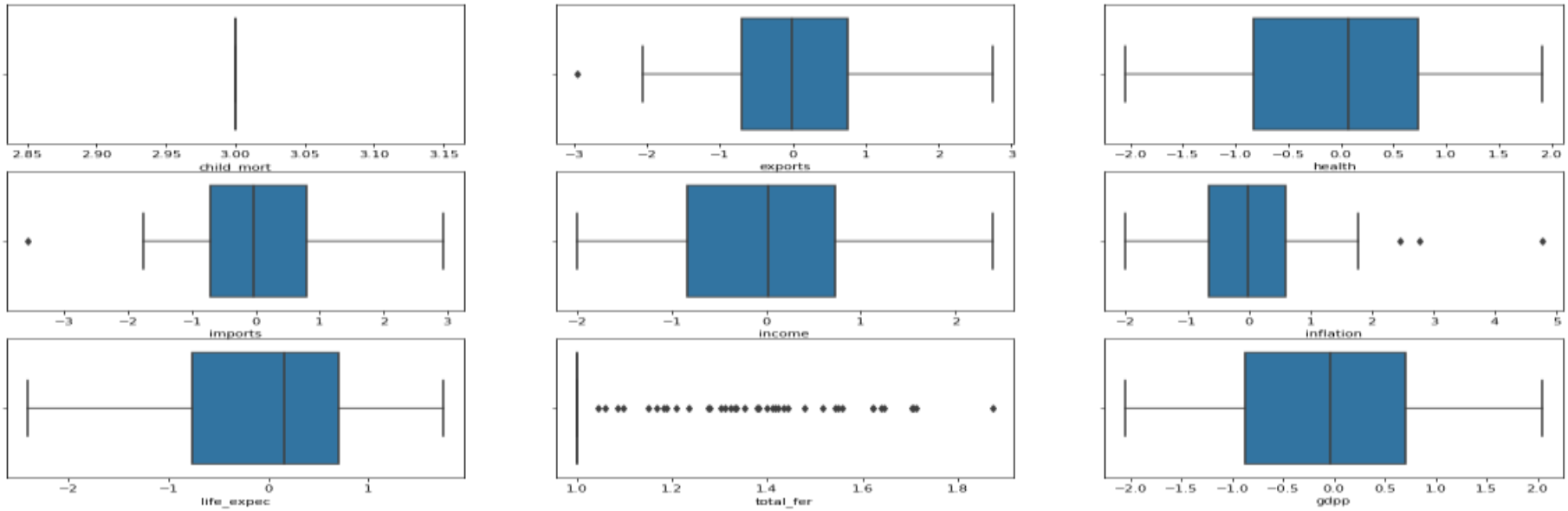


International Institute of Information Technology Bangalore

upGrad

Analysis Of Outliers & Soft Capping

- ❑ **There seems to be outliers in every single variable. This is a very delicate situation in terms of Business problem statement & Clustering analysis.**
- ❑ **If we apply outlier treatment by Deletion based on IQR values, this will remove few countries from the list that would have really deserved the Financial Aid.**
- ❑ **If we do not apply Outlier treatment, it can impact the clustering model, as the presence of Outlier can change the centroid (K-Means) of the cluster.**
- ❑ **Thus, we have used Soft Capping.**

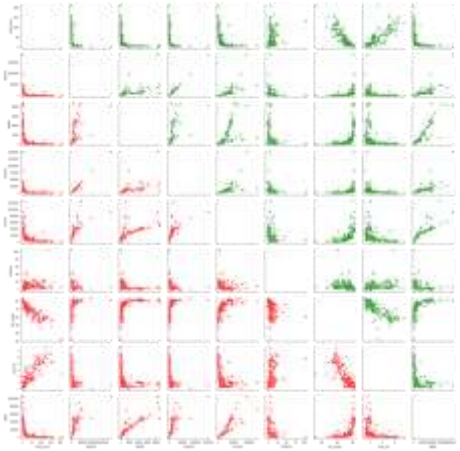


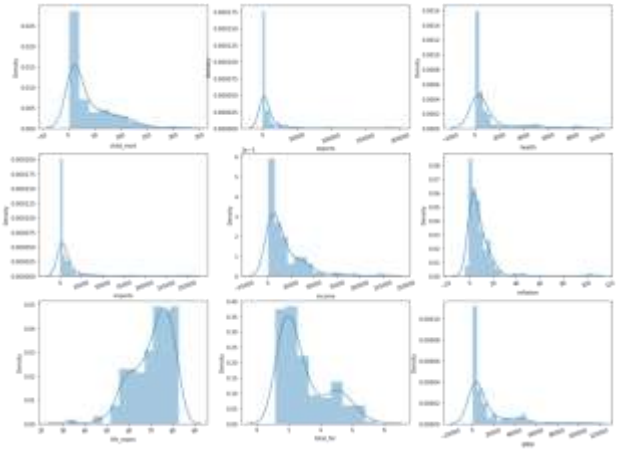
International Institute of Information Technology Bangalore

upGrad

Exploratory Data Analysis(EDA)

- ❑ Most of the data point are 'NOT Normally' distributed. Their variance are also different. Their range is also different.
- ❑ All the above points indicates the need of standardizing the data before we build the model. Since we need to compute the Euclidean distance between the data points, it is important to ensure that the attributes with a larger range of values do not out-weight the attributes with smaller range. Thus, scaling down of all attributes to the same normal scale is important here.
- ❑ Distribution plots are very important. We can get a rough idea of the no. of clusters from the plots peaks.
- ❑ Almost all the plots have more than one peaks. Like child_mort, income, export, gdpp plots are having more than 2 peaks which clearly says that there can be more than two clusters into which we can categorize the countries.
- ❑ Their ranges are also different. All the above points indicates the need of standardizing the data before we build the model.
- ❑ Since we need to compute the Euclidean distance between the data points, it is important to ensure that the attributes with a larger range of values do not out-weight the attributes with smaller range. Thus, scaling down of all attributes to the same normal scale is important here.
- ❑It can be observed from the pairplot and heatmap that there are high correlations between some variables but it will not affect on clustering.





International Institute of Information Technology Bangalore

upGrad

Hopkins Method

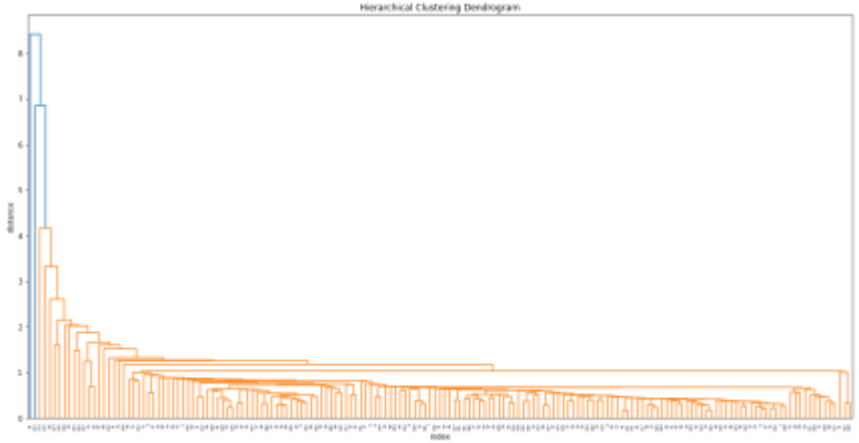
- Before we apply any clustering algorithm to the data, it's important to check whether the given data has some meaningful clusters or not. This in general means the given data is not random. The process to evaluate the data to check if the data is feasible for clustering or not is known as the clustering tendency. To check cluster tendency, we use Hopkins test.
- Hopkins test examines whether data points differ significantly from uniformly distributed data in the multidimensional space.
- Hopkins Statistic over .70 is a good score which says that the data is good for cluster analysis.
- A 'Hopkins Statistic' value close to 1 indicates that the data is highly clustered, random data will tend to result in values around 0.5, and uniformly distributed data will tend to result in values close to 0.
- Use the Hopkins Statistic function by passing the above data frame as a parameter hopkins(df_scaled) → 0.937892432284877 2

International Institute of Information Technology Bangalore

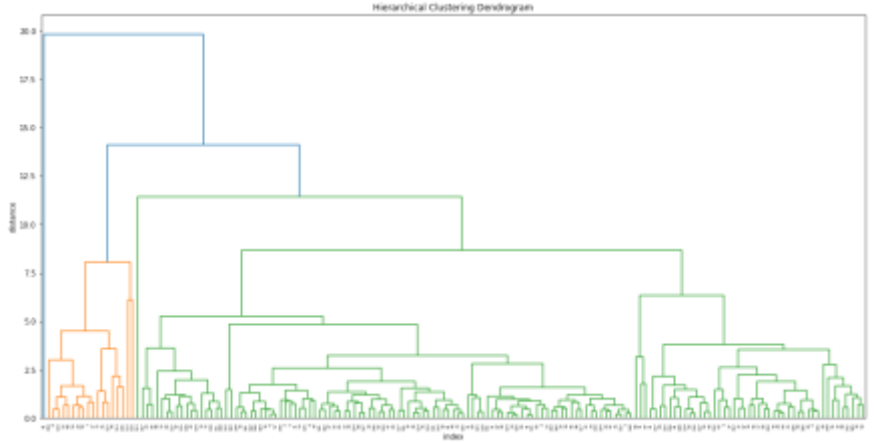
upGrad

Using Hierarchical clustering to identify the optimal cluster value

- **We will use Hierarchical Clustering to identify appropriate cluster size with a good split of data. (Max Intra-Cluster distance & Min Inter-Cluster Distance)**
- **Single linkage: Here, the distance between 2 clusters is defined as the shortest distance between points in the two clusters.**
- **Complete linkage : Here, the distance between 2 clusters is defined as the maximum distance between any 2 points in the clusters.**
- **From the below Dendrograms, it is evident that 'Complete Linkage' give a better cluster formation. So we will use Complete linkage output for our further analysis. We will build two iterations of clustering with 3 & 4 clusters (based on inputs from the above Dendrogram with Complete Linkage) and analyze the output.**



Single Linkage

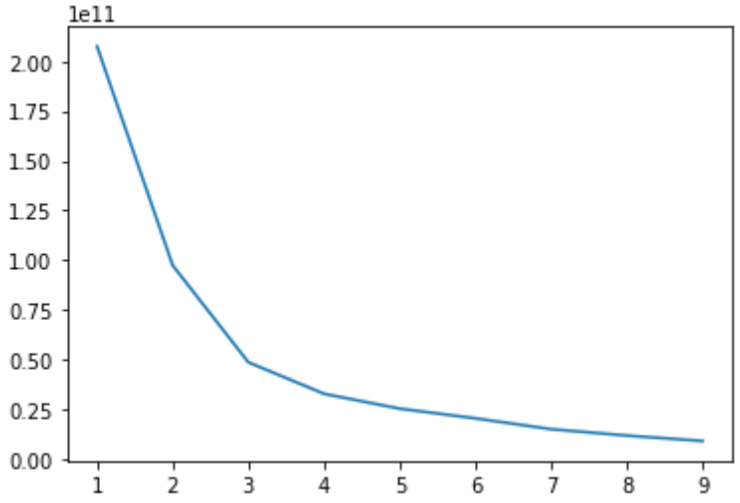


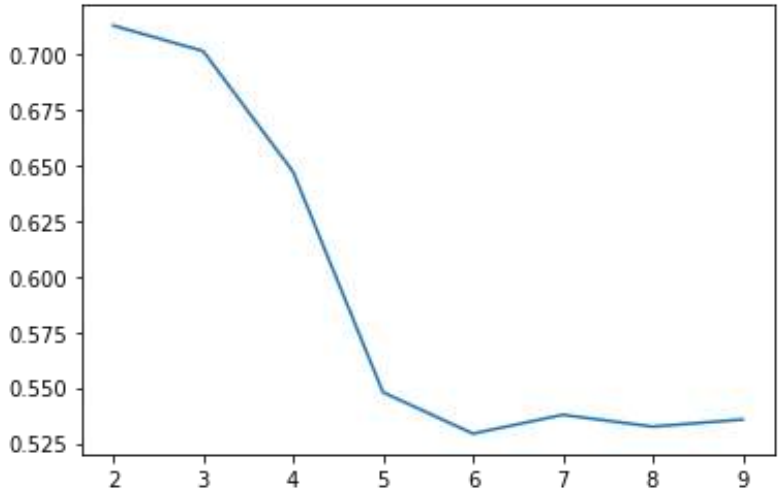
Complete Linkage

International Institute of Information Technology Bangalore

upGrad

Elbow-curve/SSD & Silhouette analysis



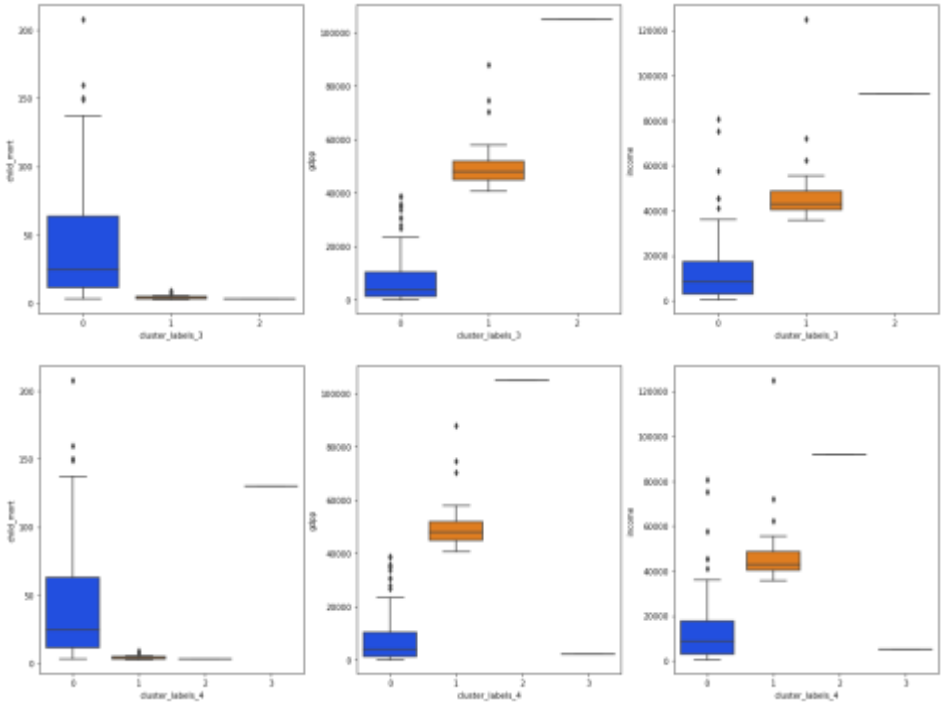


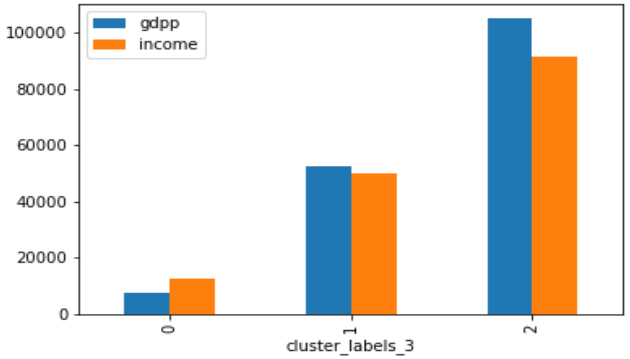
❑ **From the above validations(Elbow Curve & silhouette analysis), we could see that 3,4 or 5 clusters are optimal number of clusters to be used. We will try 3 different iterations in K-Means clustering using 3,4 and 5 clusters and analyze the results.**

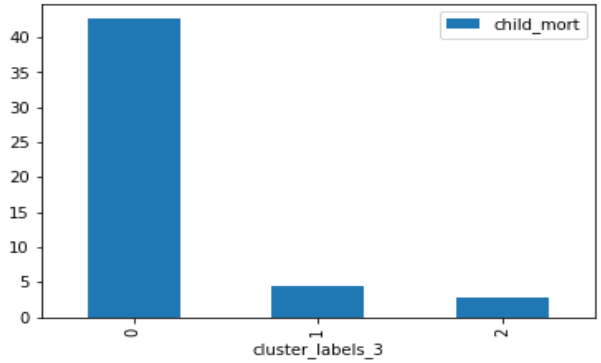
International Institute of Information Technology Bangalore

upGrad

K-Means Cluster method to build the final cluster model







[Box plot on various variable against the CLUSTER_ID to visualize the spread of the data]

❑ **Cluster 0 has the Highest average Child Mortality rate of ~42 when compared to other 3 clusters, and Lowest average GDPP & Income of ~ 7551 & 12641 respectively. All these figures clearly makes this cluster the best candidate for the financial aid from NGO. We could also see that Cluster 0 comprises of ~89% of overall data, and has ~148 observations in comparison to 167 total observations This seems to be a problem. This means that Hierarchical clustering is not giving us a good result as 89% of the data points are segmented into that cluster. We also saw that increasing the cluster number is not solving this problem. We will perform K-Means Clustering and check how that turns out to be.**

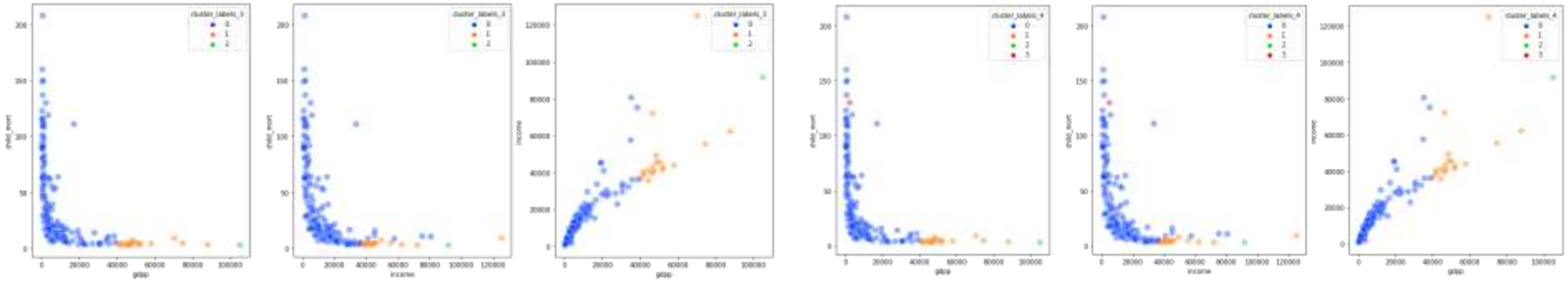
International Institute of Information Technology Bangalore

upGrad

Hierarchical Clustering

❑ **From the 2 iterations of Hierarchical Clustering, it is evident that 3 clusters is ideal number of clusters, because when we used 4 clusters, we could see that Nigeria was added as a separate segment. Since Nigera could be a possible candidate for financial aid in terms of their child mortality rate.**

❑ **Cluster 0 has the Highest average Child Mortality rate of ~42 when compared to other 3 clusters, and Lowest average GDPP & Income of ~ 7551 & 12641 respectively. All these figures clearly makes this cluster the best candidate for the financial aid from NGO. We could also see that Cluster 0 comprises of ~89% of overall data, and has ~148 observations in comparison to 167 total observations This seems to be a problem. This means that Hierarchical clustering is not giving us a good result as 89% of the data points are segmented into that cluster. We also saw that increasing the cluster number is not solving this problem. We will perform K-Means Clustering and check how that turns out to be.**

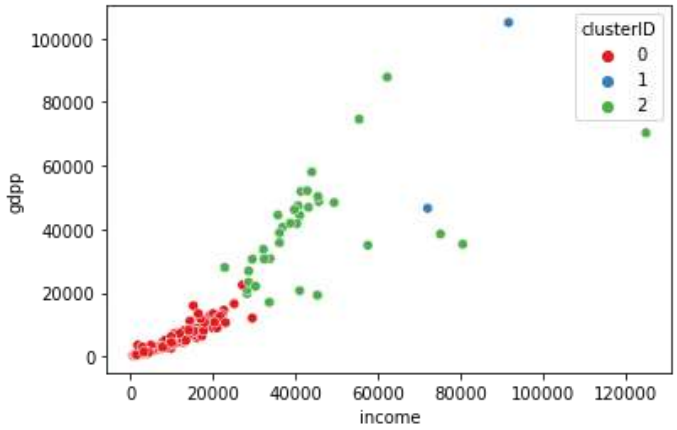


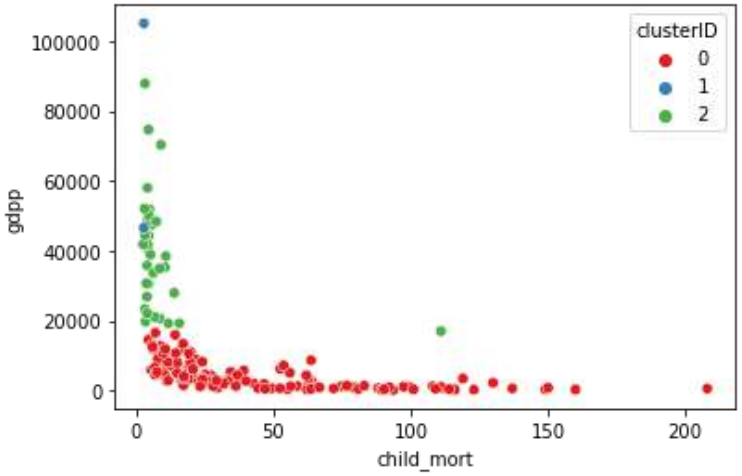
[Scatter plot on various variables to visualize the clusters based on them]

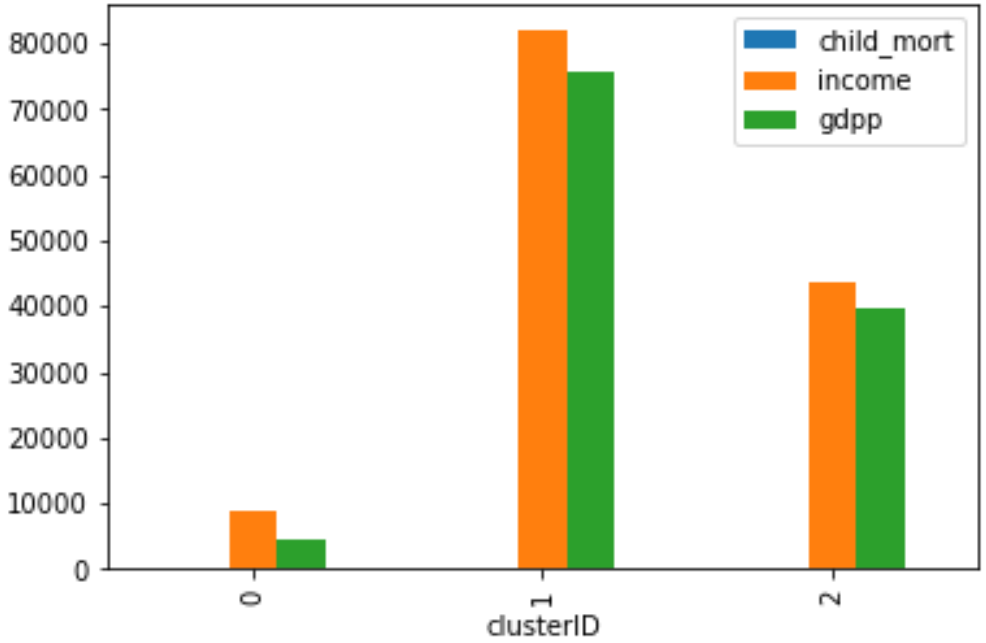
International Institute of Information Technology Bangalore

upGrad

K-Means Clustering Profiling







❑ As we can see cluster 2 has low income and gdpp we need to aid these countries.

International Institute of Information Technology Bangalore

upGrad

Final Decision Making

- Top 10 Recommended countries which are in dire need of funds:

(Top 5 marked as bold)

- **Burundi**
- **Liberia**
- **Congo, Dem. Rep.**
- **Niger**
- **Sierra Leone**
- Madagascar
- Mozambique
- Central African Republic
- Malawi
- Eritrea

International Institute of Information Technology Bangalore

upGrad

CONCLUSION & RECOMMENDATIONS

❑ All the countries have been categorized into 3 clusters : Developed, Developing and Under Developed countries from a dataset that has been provided containing 167 countries with their corresponding socio-economic and health factors.

❑ Based on our Clustering Analysis, top 10 countries from the 'Under Developed Countries' cluster has been identified and recommended which are in dire need of the Financial Aid from the Help International NGO. Recommendation has been done based on K-Means clustering with number of clusters as 3 and considering financial factor first. This output is purely based on the dataset we used and various analytical methodology we performed. These countries have: low gdpp low income and high child mortality.

International Institute of Information Technology Bangalore

upGrad

THANK YOU