CLUSTERING ASSIGNMENT REPORT

- Submitted by

Saikat Chowdhury

Assignment Objective and Methodology

❖**Business Objective**: 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. Using country socio-economic data we need to identify countries which are in the direst need of aid.

❖**Methodology**:

A nine feature socio-ecomic factor 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 eg % Health converted to Heath 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 Kmeans 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.

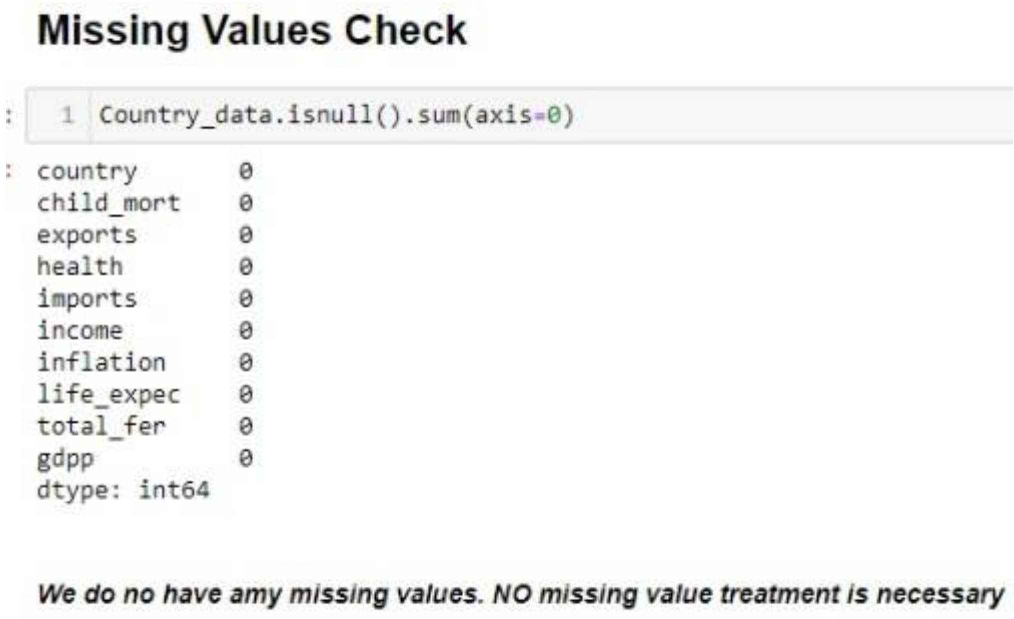
Approach of the Clustering Analysis

Based on the business problem and looking at the dataset at a high level, I will be following the below approach to solve this problem.

1. Check for missing value, and treatment
2. Check for outlier and treatment
3. Perform the basic EDA to find the variablity and distribution of the data, so as to identify if we need t scaling the data
4. Data Scaling if necessary
5. Use Hopkins Method to check if the dataset is good enough for a cluster analysis
6. Using Hierarchical clustering to identify the optimal cluster value.
7. Use Silhouette and Elbow method to validate the optimal cluster values.
8. Use K-Means Cluster method to build the final cluster model.
9. Analyse the cluster that is representing the countries that will solve the Business Problem.
10. Present the final report

Check for missing value, and treatment

1. Data frame has data about various countries and their socio-economic factors. Few are in % and others in absolute values.
2. Data frame has 10 Columns and 167 Rows
3. One variable is 'Object' Type, and rest all are 'Int' or 'Float' type
4. Descriptive Statistics tells us that there is variablity in the data, and will require scaling before model building.



Missing Values Check

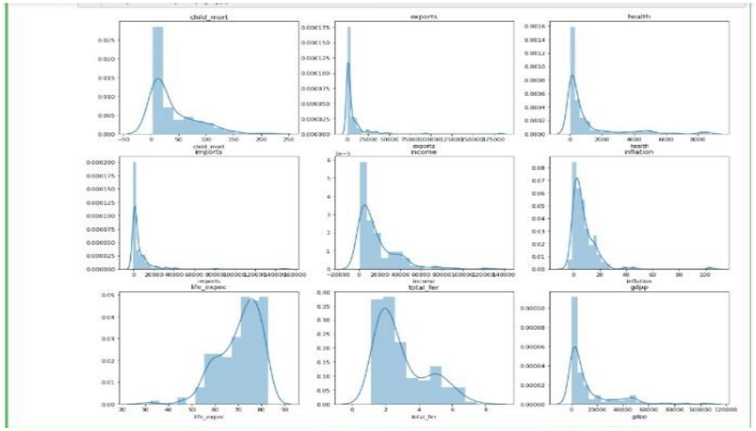
```
Country_data.isnull().sum(axis=0)
```

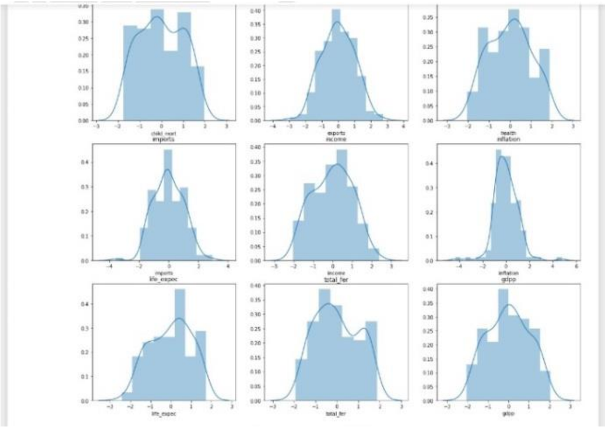
```
country       0
child_mort    0
exports       0
health        0
imports       0
income        0
inflation     0
life_expec    0
total_fer     0
gdpp          0
dtype: int64
```

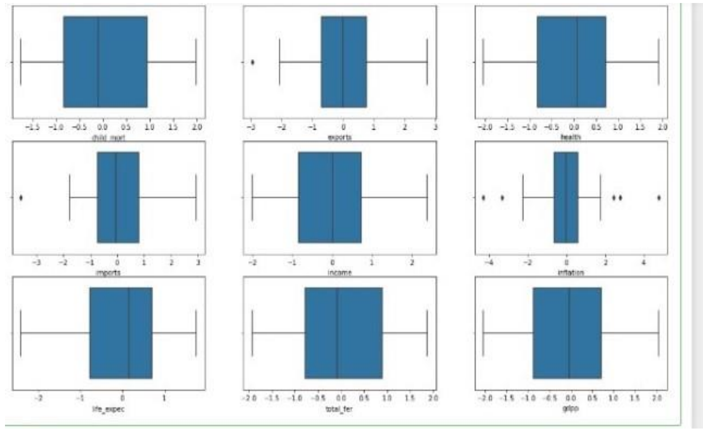
We do no have amy missing values. NO missing value treatment is necessary

Check for outlier and treatment

- Plotting all the features to visualize and look their distributions.
- 'child_mort', 'exports', 'health', 'imports', 'income', 'inflation', 'life_expec', 'total_fer', 'gdpp'





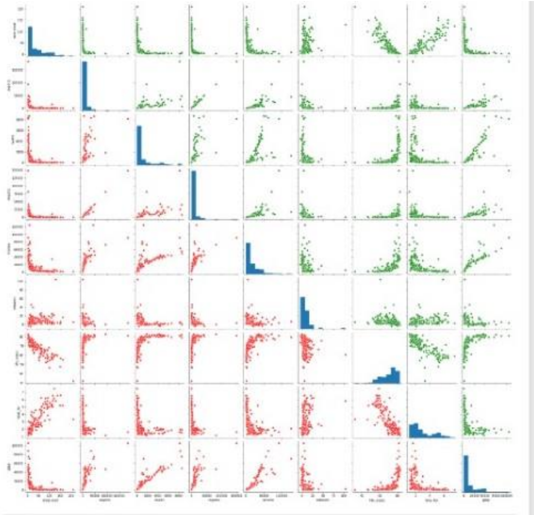


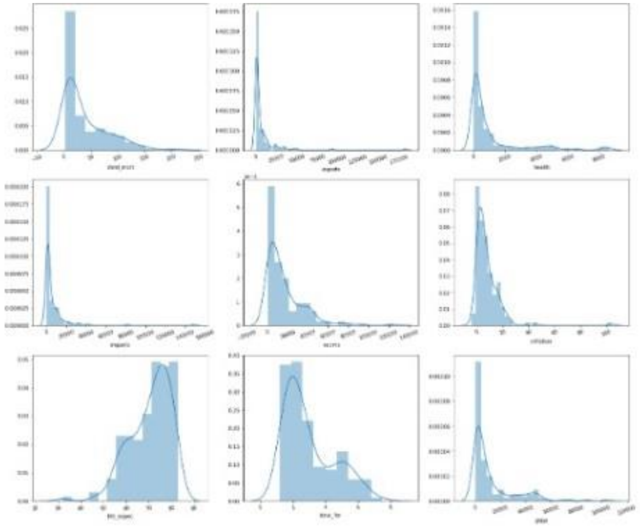
Outlier Analysis Insights

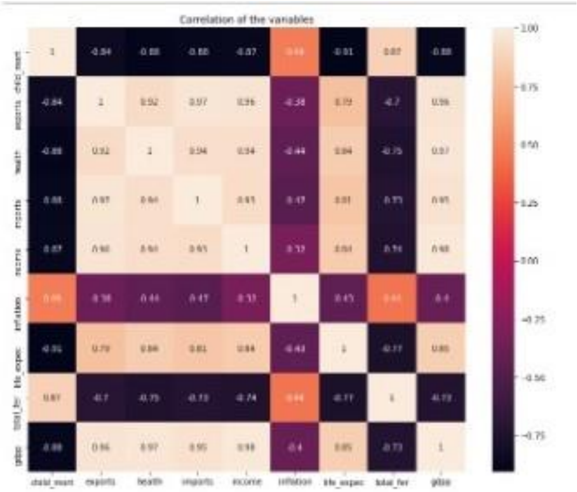
- There seems to be outliers in every single variable. This is a very delicate situation in terms of Business problem statement & Clustering analysis.
- If we apply outlier treatment by Deletion based on IQR values, this will remove few countries from the list that would have really deserved the Financial 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.
- After considering all these scenarios and the business call, I have decided to use **SOFT CAPPING** (less number of observations; 167):
  - to the lower range outliers for 'child_mort','inflation','total_fer' as the values of these variables need to be high to be eligible for financial aid and
  - to the upper range outliers to the rest of the variables as values of these variables for the countries need to be less to be eligible for financial aid from the NGO.

Perform the basic EDA to find the variablity and distribution of the data

- Most of the data point are 'NOT Normally' distributed.
- Their variance are also different.
- Their range 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.





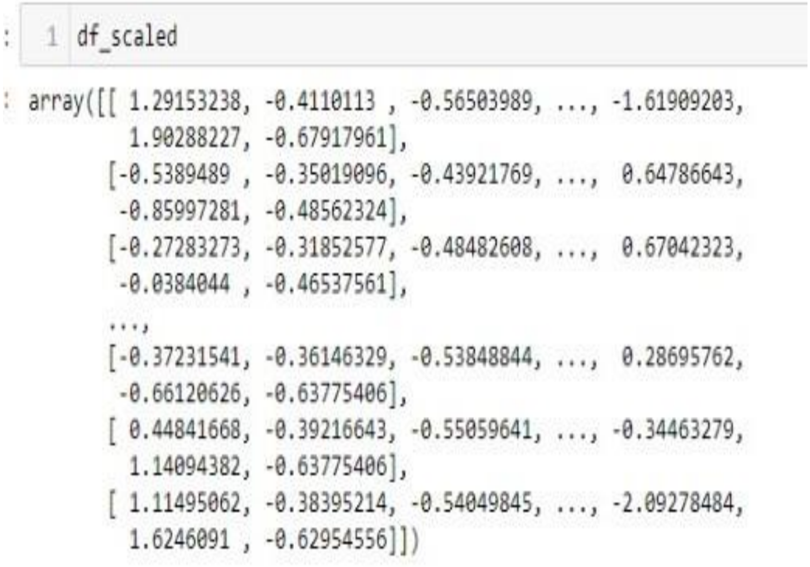


Insights for EDA Analysis

- Distribution plots are very important. We can get a rough idea of the no. of clusters from the plots peaks.
- Most of the data point are 'Not Normally' distributed.
- 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 differnt. All the above points indicates the need of standardising 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.

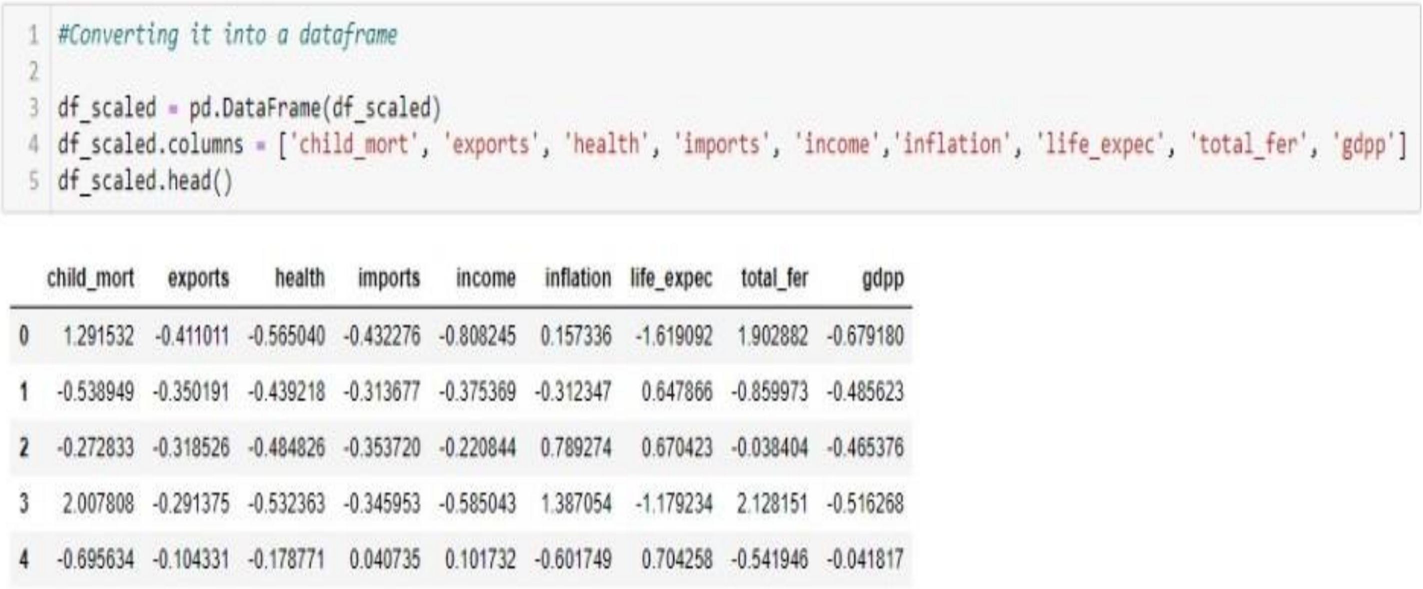
Scaling the Data

- We will use Standardization method for scaling the data



```
df_scaled
```

```
array([[ 1.29153238, -0.4110113 , -0.56503989, ..., -1.61909203,
         1.90288227, -0.67917961],
       [-0.5389489 , -0.35019096, -0.43921769, ...,  0.64786643,
        -0.85997281, -0.48562324],
       [-0.27283273, -0.31852577, -0.48482608, ...,  0.67042323,
        -0.0384044 , -0.46537561],
       ...,
       [-0.37231541, -0.36146329, -0.53848844, ...,  0.28695762,
        -0.66120626, -0.63775406],
       [ 0.44841668, -0.39216643, -0.55059641, ..., -0.34463279,
         1.14094382, -0.63775406],
       [ 1.11495062, -0.38395214, -0.54049845, ..., -2.09278484,
         1.6246091 , -0.62954556]])
```



```
#Converting it into a dataframe

df_scaled = pd.DataFrame(df_scaled)
df_scaled.columns = ['child_mort', 'exports', 'health', 'imports', 'income','inflation', 'life_expec', 'total_fer', 'gdpp']
df_scaled.head()
```

	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
0	1.291532	-0.411011	-0.565040	-0.432276	-0.808245	0.157336	-1.619092	1.902882	-0.679180
1	-0.538949	-0.350191	-0.439218	-0.313677	-0.375369	-0.312347	0.647866	-0.859973	-0.485623
2	-0.272833	-0.318526	-0.484826	-0.353720	-0.220844	0.789274	0.670423	-0.038404	-0.465376
3	2.007808	-0.291375	-0.532363	-0.345953	-0.585043	1.387054	-1.179234	2.128151	-0.516268
4	-0.695634	-0.104331	-0.178771	0.040735	0.101732	-0.601749	0.704258	-0.541946	-0.041817

Use Hopkins Method

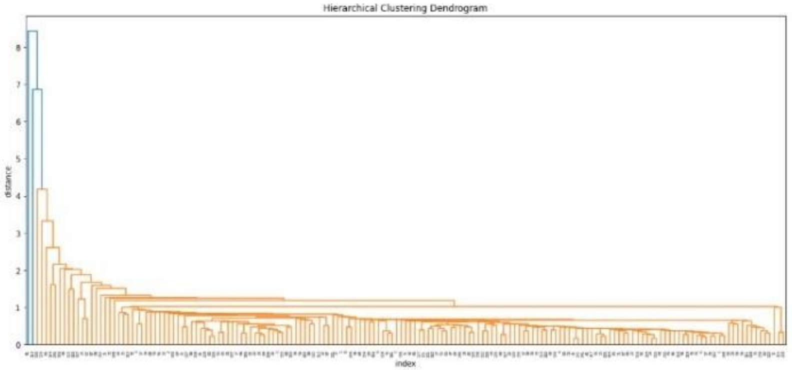
- Before we apply any clustering algorithm to the data, it's important to check whether the given data has some meaningful clusters or not. This in general means the given data is not random. The process to evaluate the data to check if the data is feasible for clustering or not is known as the clustering 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.

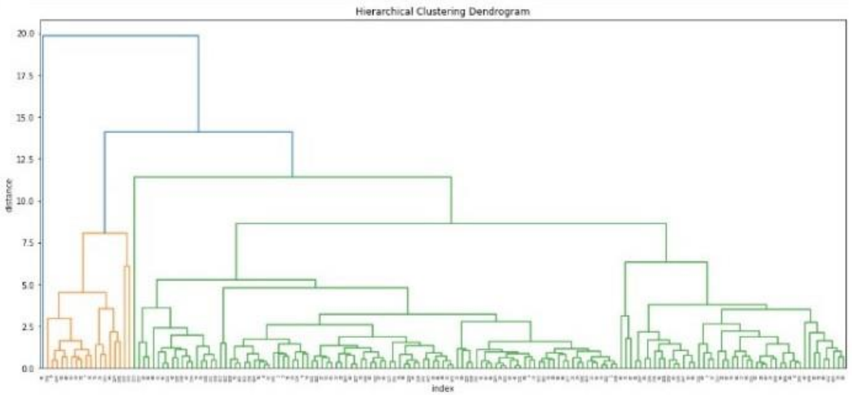
- Hopkins Statistic over .70 is a good score which says that the data is good for cluster analysis.
- A 'Hopkins Statistic' value close to 1 indicates that the data is highly clustered, random data will tend to result in values around 0.5, and uniformly distributed data will tend to result in values close to 0.

- #Use the Hopkins Statistic function by passing the above dataframe as a paramter
- hopkins(df_scaled)
- -> 0.9415971161683127

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.





- Single Linkage

Complete Linkage

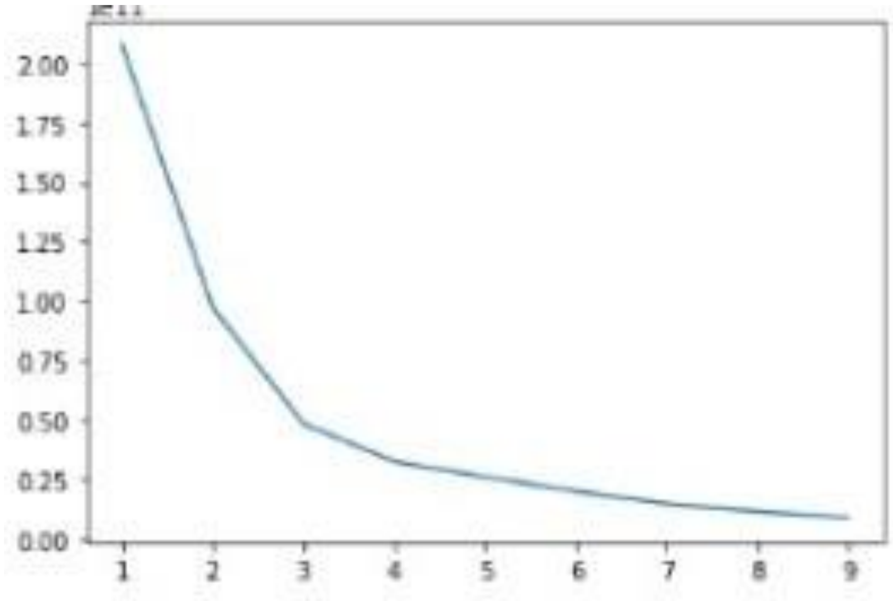
- From the above Dendrograms, it is evident that 'Complete Linkage' give a better cluster formation. So we will use Complete linkage output for our further analysis. We will build two iterations of clustering with 3 & 4 clusters (based on inputs from the above Dendrogram with Complete Linkage) and analyse the output.

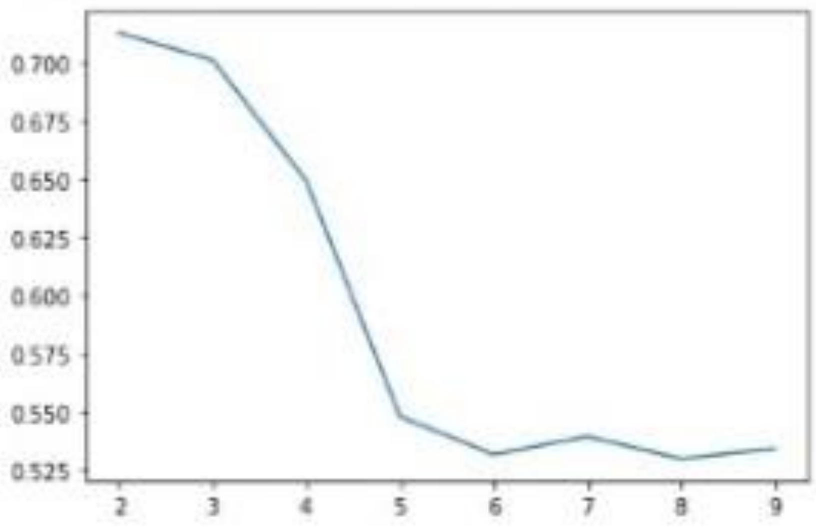
Silhouette and Elbow method

- **silhouette score=p−q/max(p,q)**
- p is the mean distance to the points in the nearest cluster that the data point is not a part of
- q is the mean intra-cluster distance to all the points in its own cluster.
- The value of the silhouette score range lies between -1 to 1.
- A score closer to 1 indicates that the data point is very similar to other data points in the cluster,
- A score closer to -1 indicates that the data point is not similar to the data points in its cluster.

- The **elbow method** runs **k-means** clustering on the dataset for a range of values for **k** (say from 1-10) and then for each value of **k** computes an average score for all clusters. By default, the distortion score is computed, the sum of square distances from each point to its assigned center.

Elbow-curve/SSD & silhouette analysis

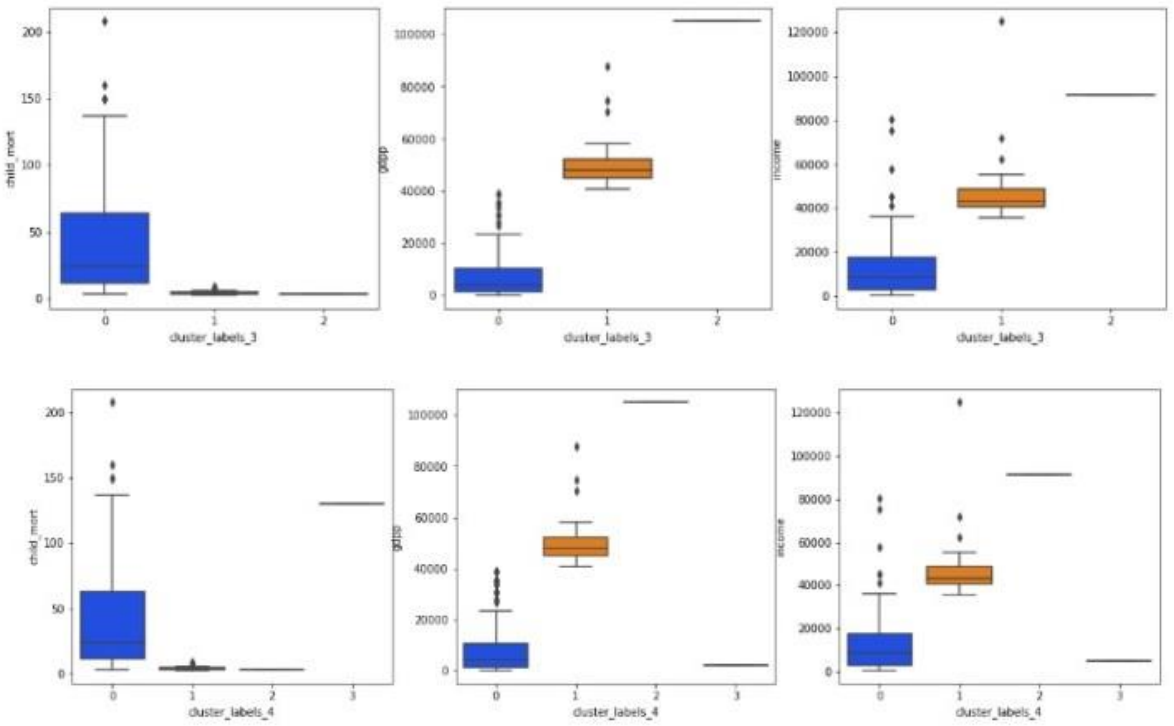


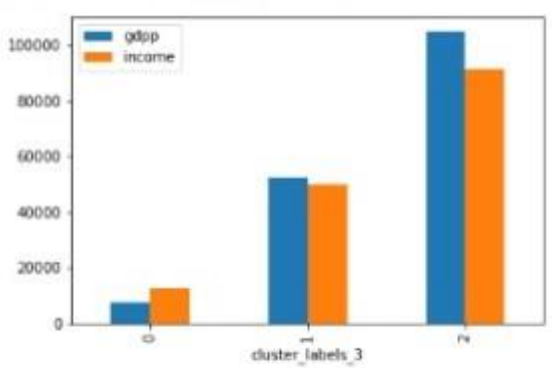


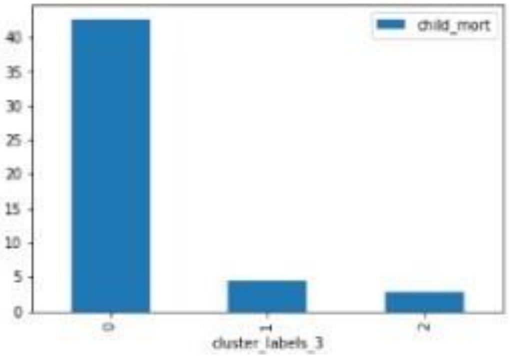
From the above validations(Elbow Curve & silhouette analysis), we could see that 3,4 or 5 clusters are optimal number of clusters to be used. We will try 3 different iterations in K-Means clustering using 3,4 and 5 Clusters and analyse the results.

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

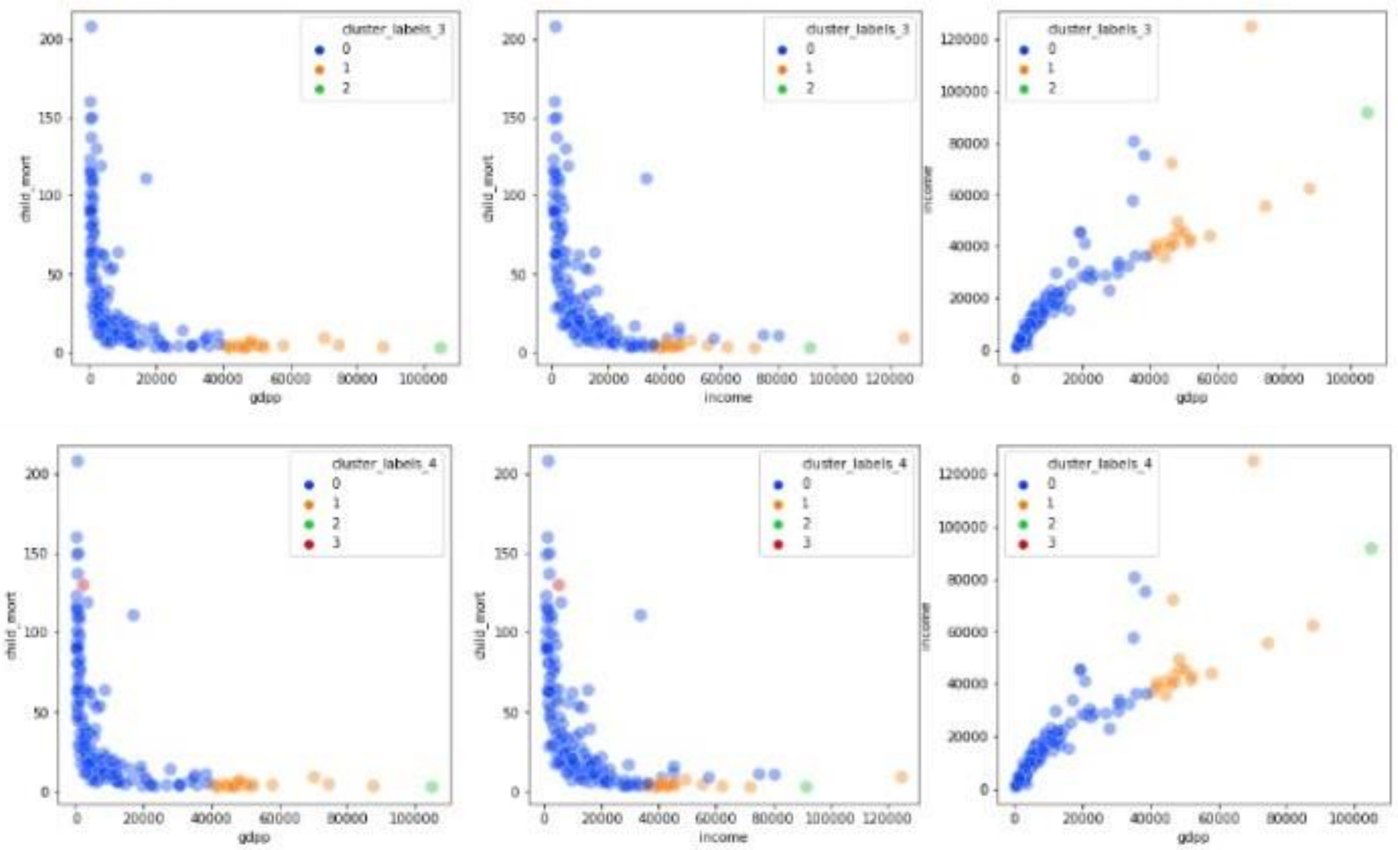






Hierarchical Clustering

- # Scatter plot on various variables to visualize the clusters based on them



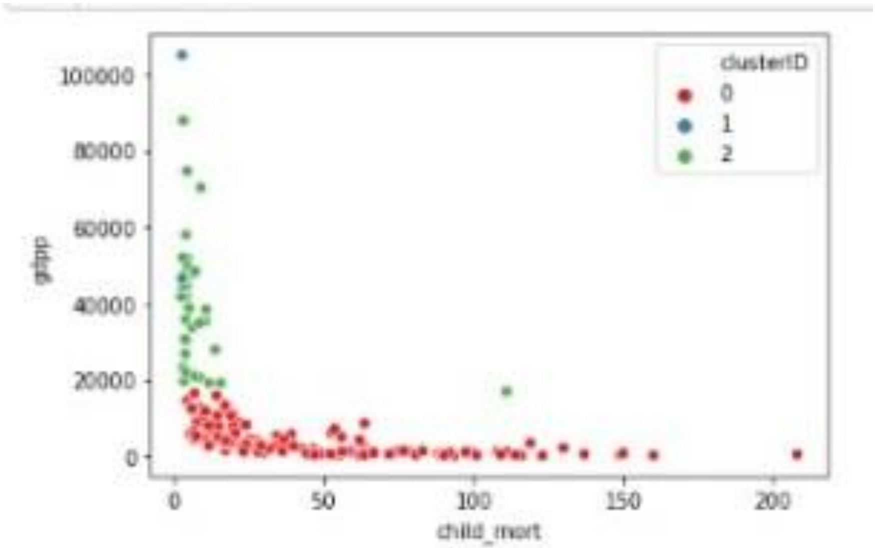
Hierarchical Clustering Insights

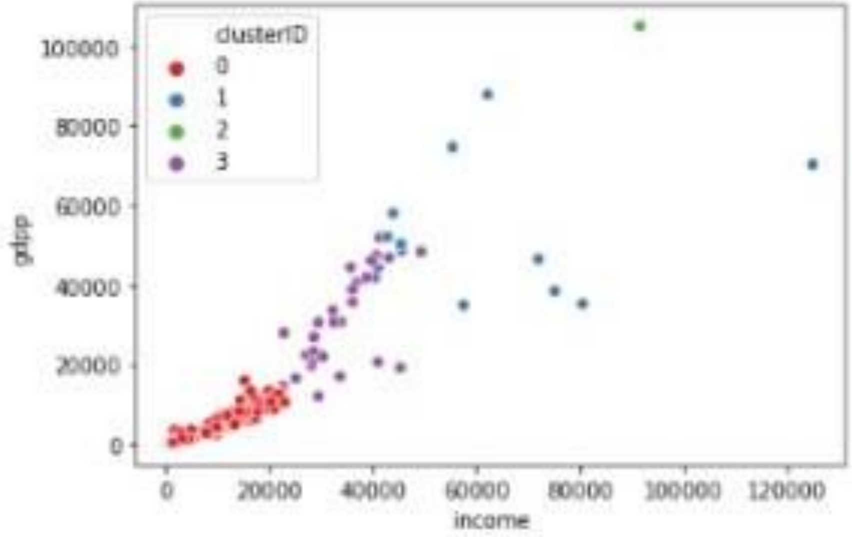
- From the 2 iterations of Hierarchical Clustering, it is evident that 3 CLUSTERS is ideal number of clusters, because when we used 4 clusters, we could see that Nigeria was added as a seperate segment. Since Nigera could be a possible candidate for financial aid in terms of their child mortality rate.

- 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 comparision 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.

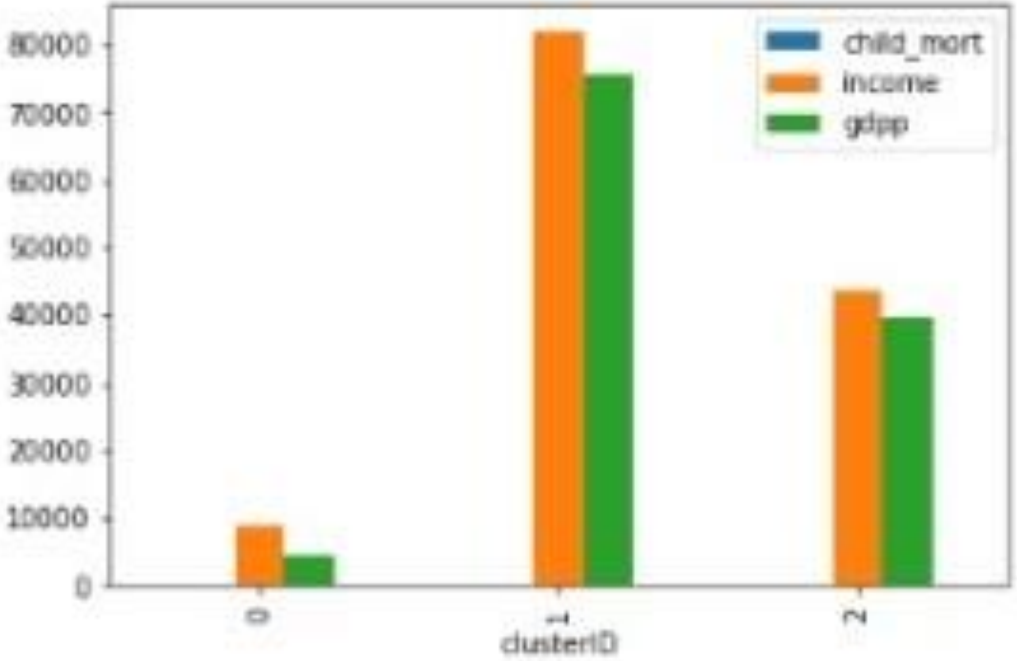
K-Means Clustering

- # Plotting the scatter plot using the clusters obtained





K-Means Clustering Profiling



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

Decision Making on the final approach

- **Based on the K-Means clustering analysis, below are the top 5 countries that are need of direct aid.**

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

Conclusion

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