CLUSTERING ASSIGNMENT

- CHETHAN BR (chethanbr86@gmail.com)

Problem Statement:

- The international humanitarian NGO called HELP 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.

- The CEO of the NGO needs a strategical and effective way to make use of 10 million dollars raised during funding programs and wants to choose which of the countries are in dire need of help.

- I hereby am categorizing countries based on socio-economic and health factors which shows overall development of the country.

- By running clustering algorithm on the data set and considering the above said factors, I am giving the names of 5 countries which need help which are detailed in upcoming pages.

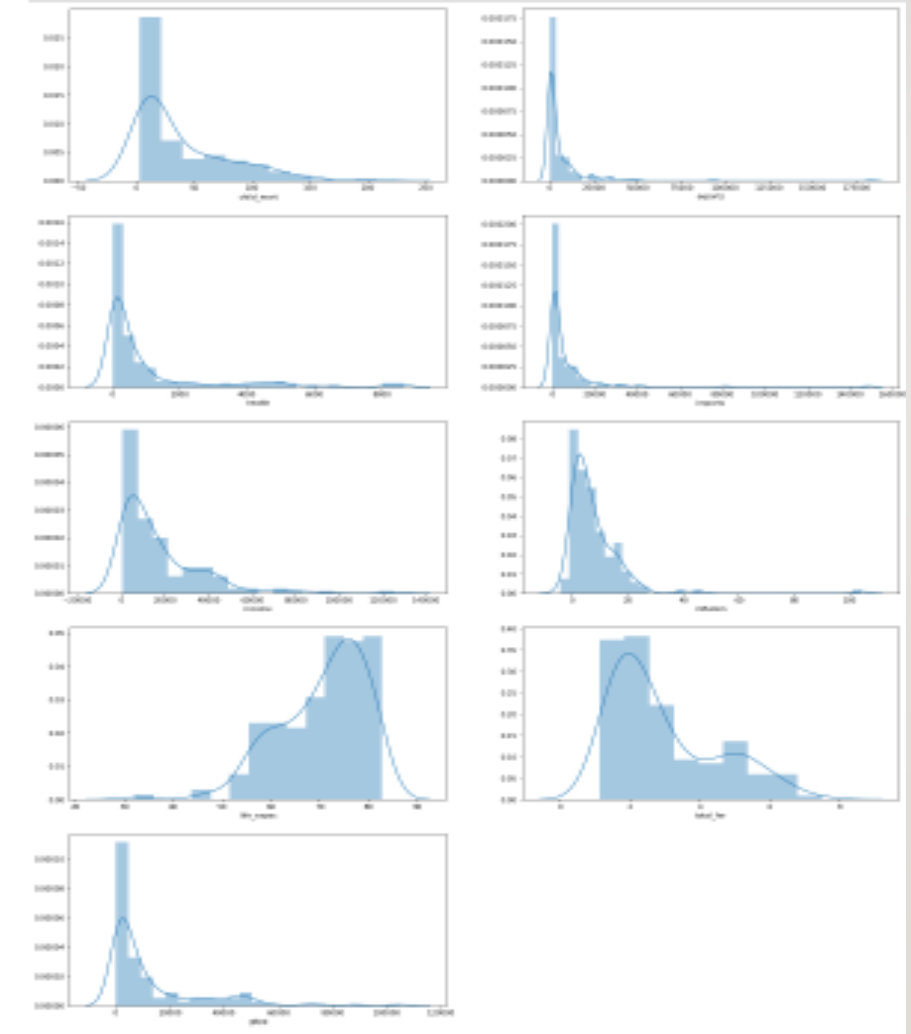
Analysis Approach

- Important libraries required for clustering like NumPy, Pandas, SKLearn which includes StandardScaler, KMeans, silhouette and SciPy libraries for hierarchy clustering including linkage, dendrogram and cut_tree are imported.
- The input dataset is imported and checked for null values, duplicate values and data type of columns.
- Certain columns are given as percentages and must be converted back to normal values. These columns include exports, health and imports. Each column values are multiplied by gdpp values and divided by 100 to get normal values.

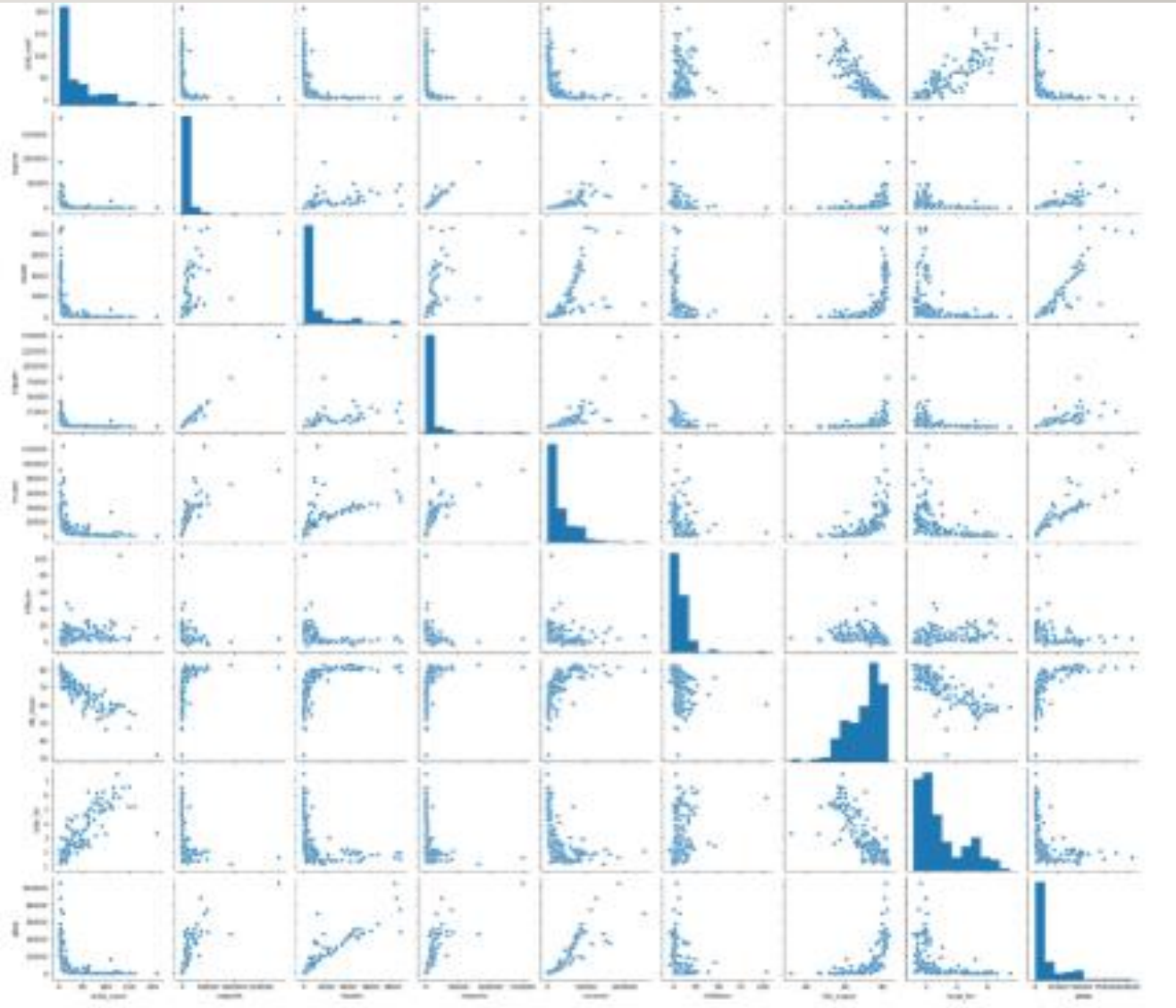
```
help_NGO_df['exports'] = help_NGO_df['exports']*help_NGO_df['gdpp']/100
help_NGO_df['health'] = help_NGO_df['health']*help_NGO_df['gdpp']/100
help_NGO_df['imports'] = help_NGO_df['imports']*help_NGO_df['gdpp']/100
```

- Univariate Analysis is done by plotting a distplot where we can see how the data points are distributed.
- Bivariate analysis is done by plotting a pair plot and heatmap.

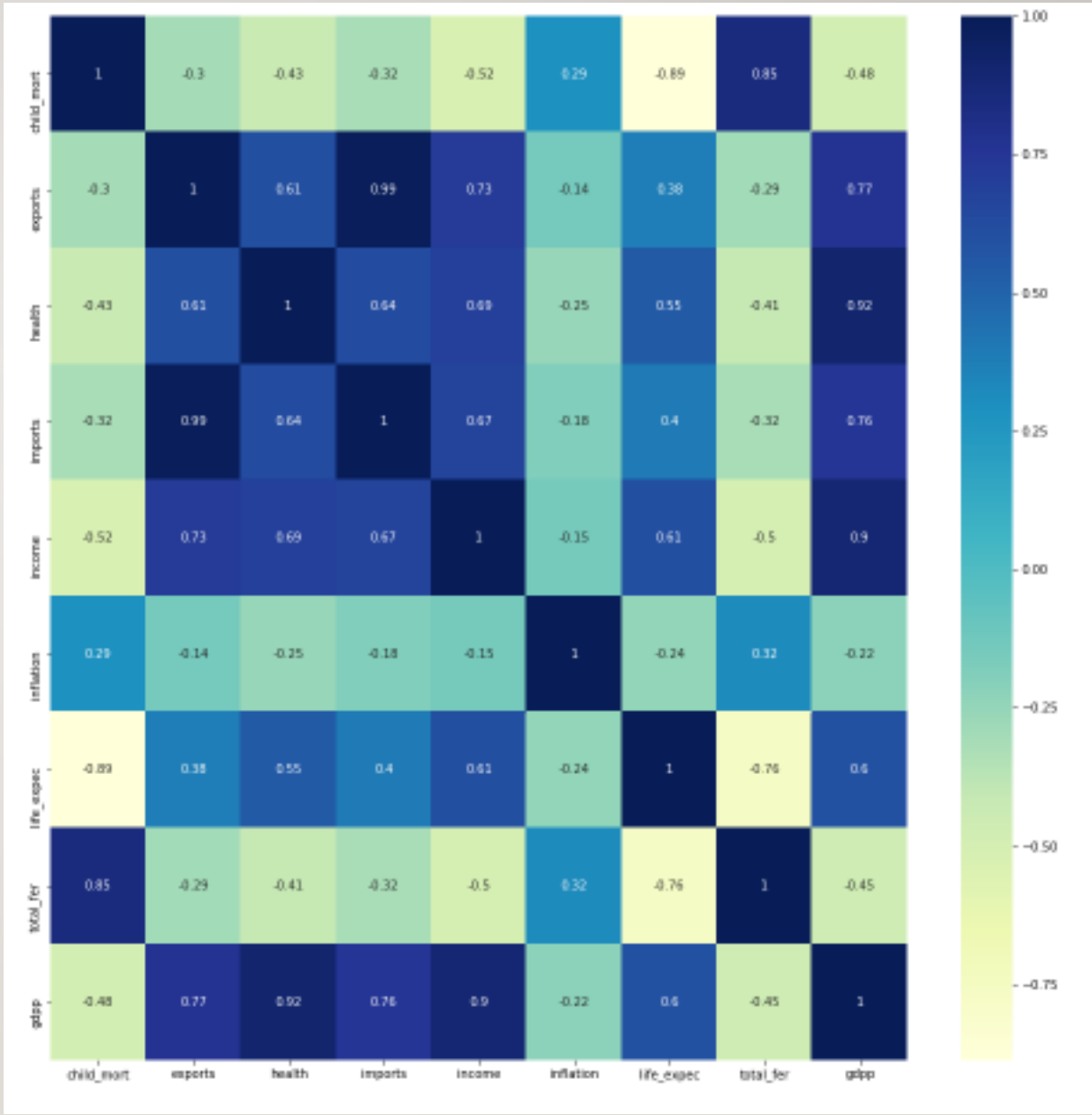
- Dist. Plot showing how normally distributed the data points in the data set.



- Pair plot shows how each variable is related with other.



- The heatmap shows how each variable is correlated with each other.
- We can clearly see that:
  - child_mort has high negative correlation with life_expec
  - gdpp has high positive correlation with health, exports, imports and income
  - exports and imports are highly correlated
  - total_fer is highly positively correlated with child_mort and negatively correlated with life_expec

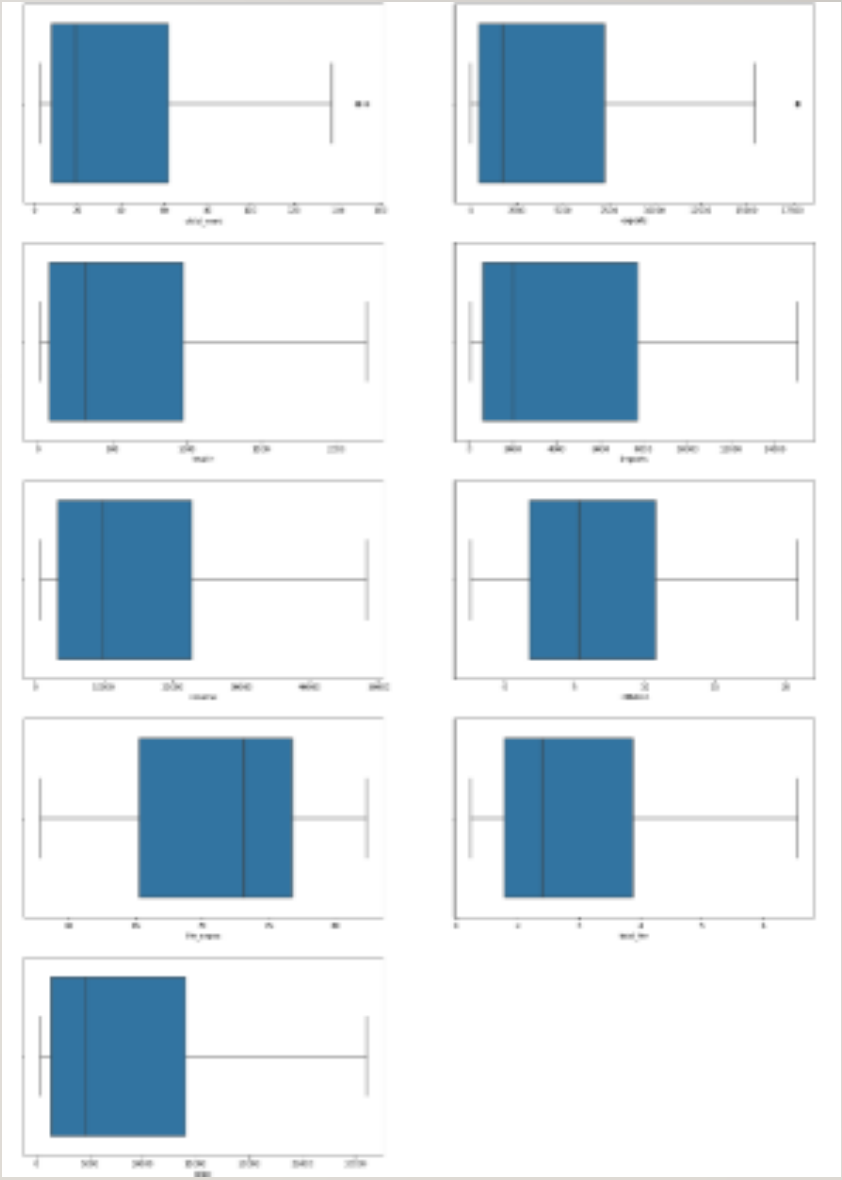


FINDING OUTLIERS

Each column might have a value which is very much outside a certain range of values and these values are called as outliers.

If we do not treat the outliers, the dataset which is subjected to any machine learning algorithm might give different outcome .

We can either delete or cap the outliers and with former we might lose important data and hence the variables with outliers have been capped with certain upper and lower limits suitable for the dataset.



Scaling of variables

Once the outliers are treated, we can move to the next step which is scaling the variables so that all the variables are treated within the same limit i.e; Standard Scaler will normalize all column values in such a way that the variables will have mean as 0 and standard deviation as 1. All values lie between zero and one.

Standardization:

$$z = \frac{x-\mu}{\sigma}$$

with mean:

$$\mu = \frac{1}{N}\sum_{i=1}^{N}(x_i)$$

and standard deviation:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2}$$

Hopkins Score

The Hopkins statistic is a way of measuring cluster tendency of a data set. It acts as statistical hypothesis test where null hypothesis is that the data is generated by a Poisson point process and are thus uniformly distributed. It gives a measure of how good the data is for clustering.

Clustering

Clustering is an unsupervised learning technique where we can find patterns based on similarities in the data.
Here the data will be grouped into different categories based on set of attributes.

How does it work

Different Clusters are formed based on how similar the data points in same cluster should be, known as intra-cluster distance and how far away are points so that they belong to different cluster known as inter-cluster distance. The distance measure that is used in K-means clustering is called the Euclidean Distance measure which is given by:

$$D = \sqrt{(X_1 - Y_1)^2 + (X_2 - Y_2)^2 + \ldots (X_n - Y_n)^2}$$

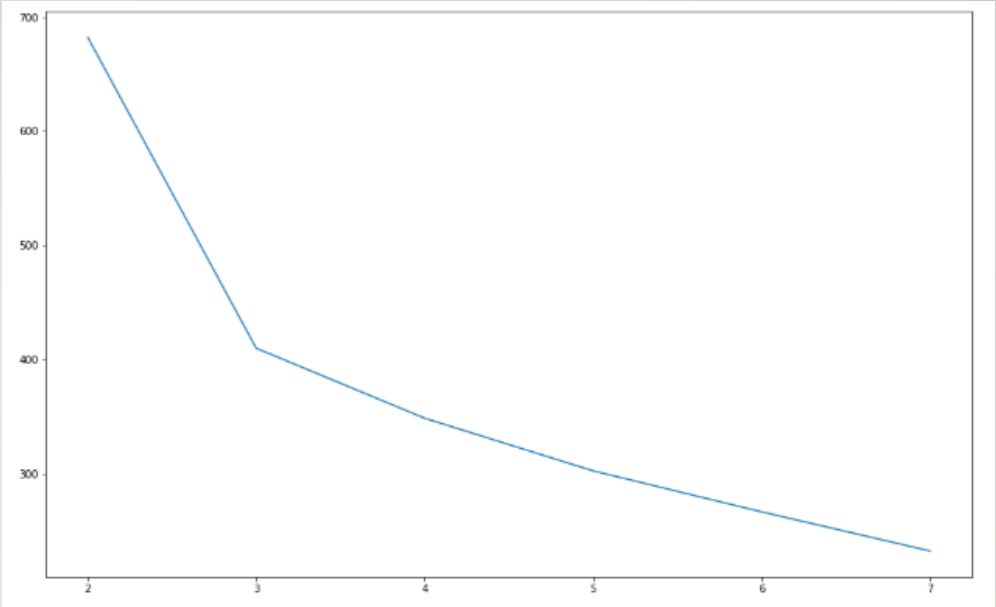
K-Means algorithm uses concept of centroid which is the arithmetic mean of all points around it.
There are 2 steps involved in Clustering called:
Assignment step where every data points to clusters are assigned depending on the distance of the points to the cluster centroid and
Optimization step where the algorithm calculates average of all points and moves the centroid until there is no change in cluster points.

Methods of Clustering

There are two methods of Clustering:

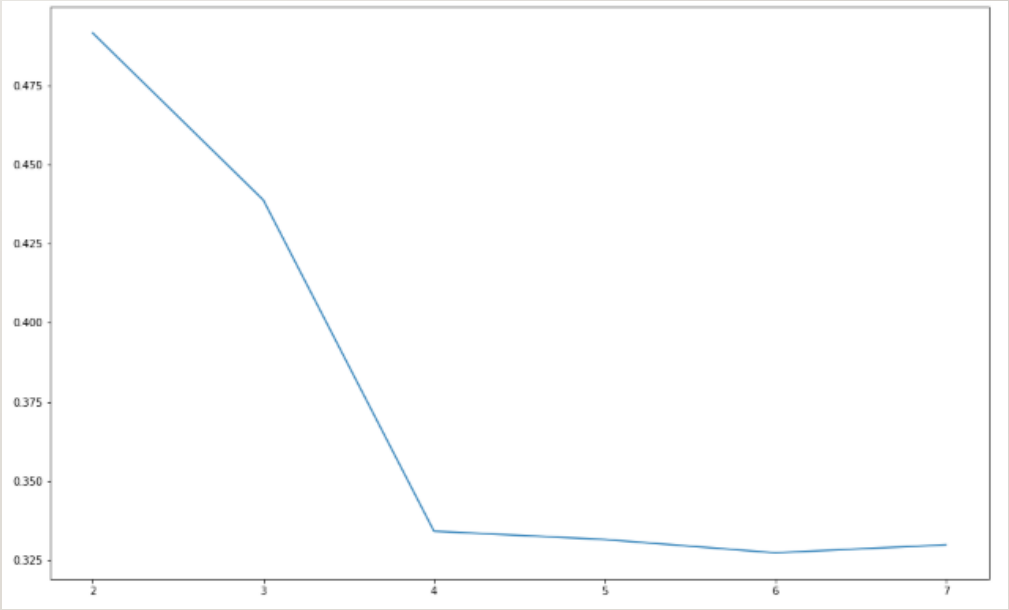
1. K-Means Clustering
    a) Elbow method
    b) Silhouette method
2. Hierarchical Clustering

Elbow method



From the graphs, I have taken the Clustering points to be 3.

Silhouette method



The clusters formed in both K-Means and Hierarchical methods are shown below.

```
help_NGO_df.labels_kmeans.value_counts()
```

```
Developing          80
Under-Developed     46
Developed           41
Name: labels_kmeans, dtype: int64
```

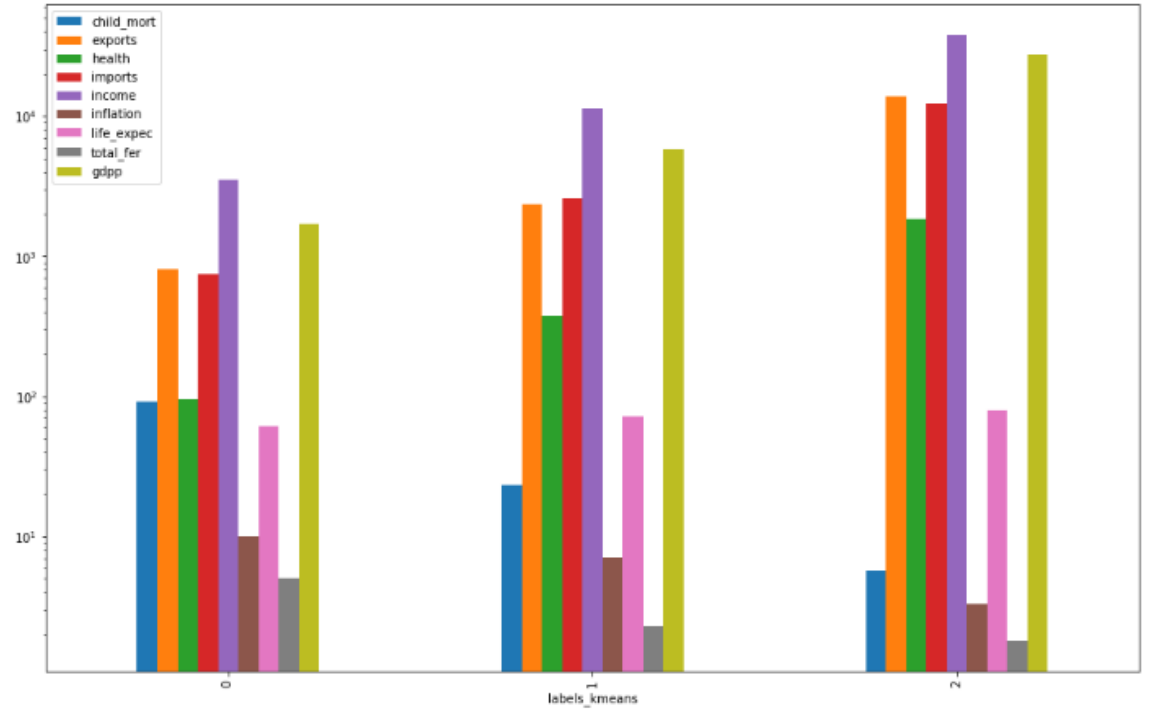
```
help_NGO_df.labels_hirarchical.value_counts()
```

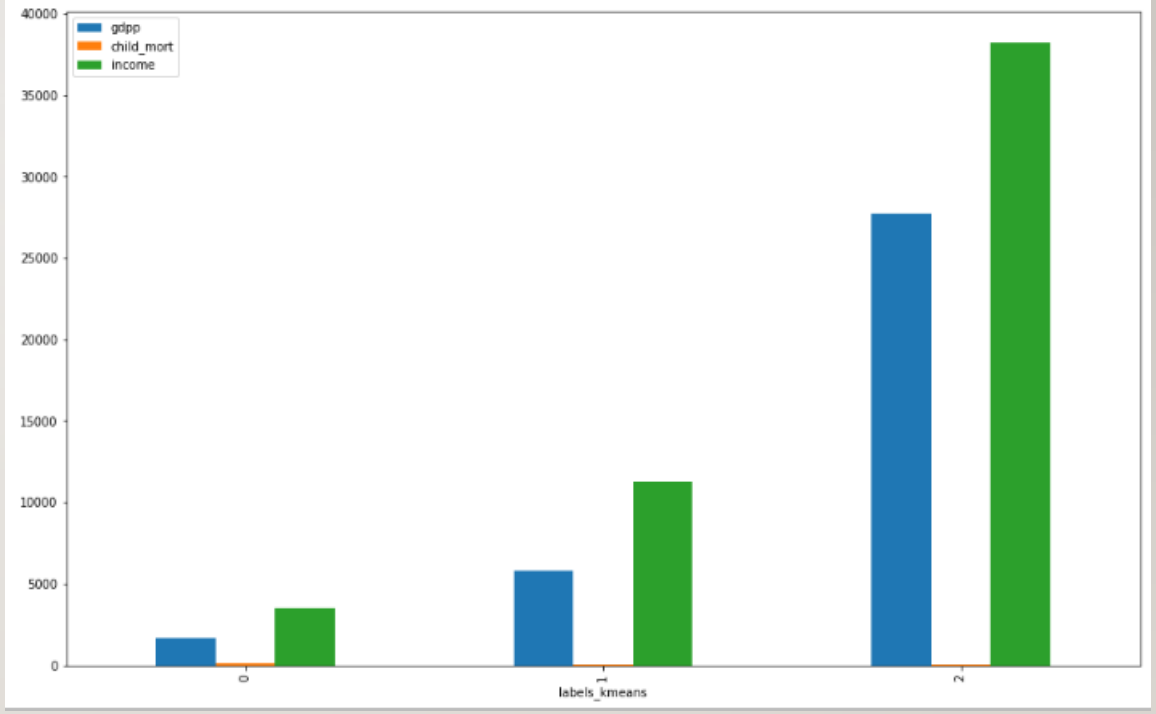
```
Developing          92
Developed           40
Under-Developed     35
Name: labels_hirarchical, dtype: int64
```

This shows our clustering through K-Means as well as Hierarchical are almost relatable .

Analysis K-Means and Hierarchical

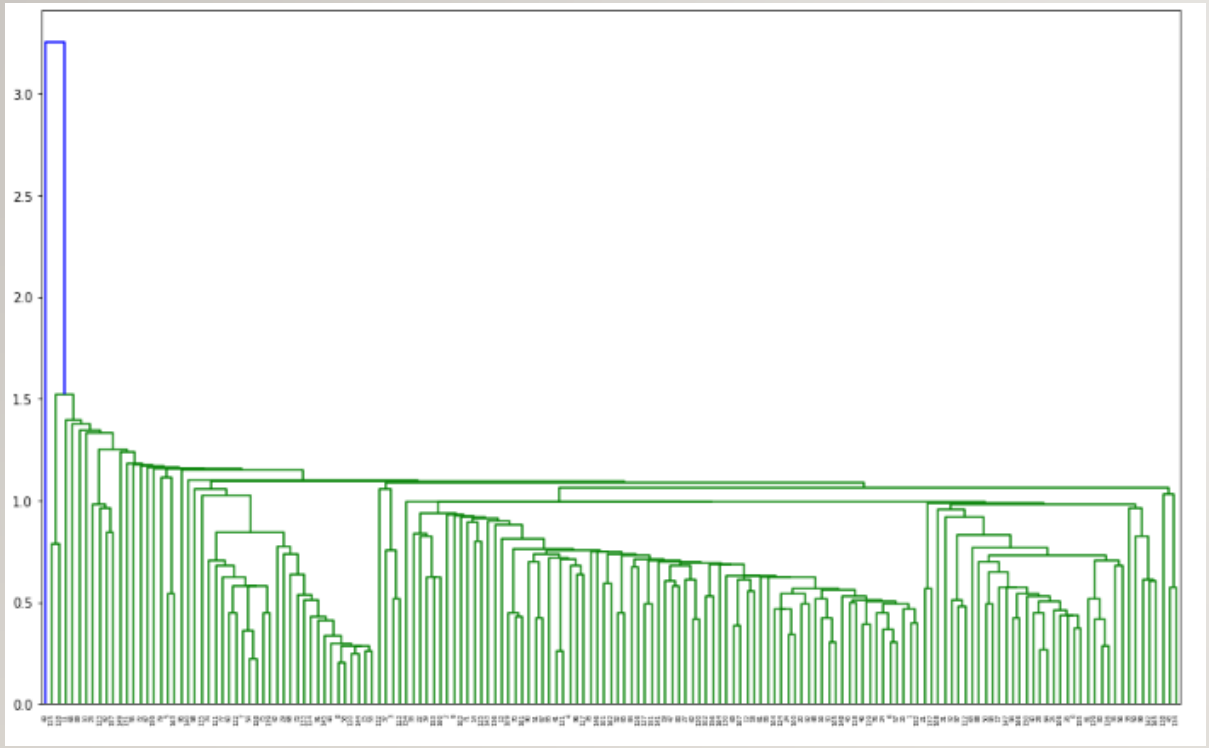
Once the number of Cluster points are obtained, Cluster labels are formed based on Euclidean measure.
The required parameters are then grouped together with labels to plot the graphs for analysis.
The boxplot shows how child_mort stacks compared to GDPP and Income.



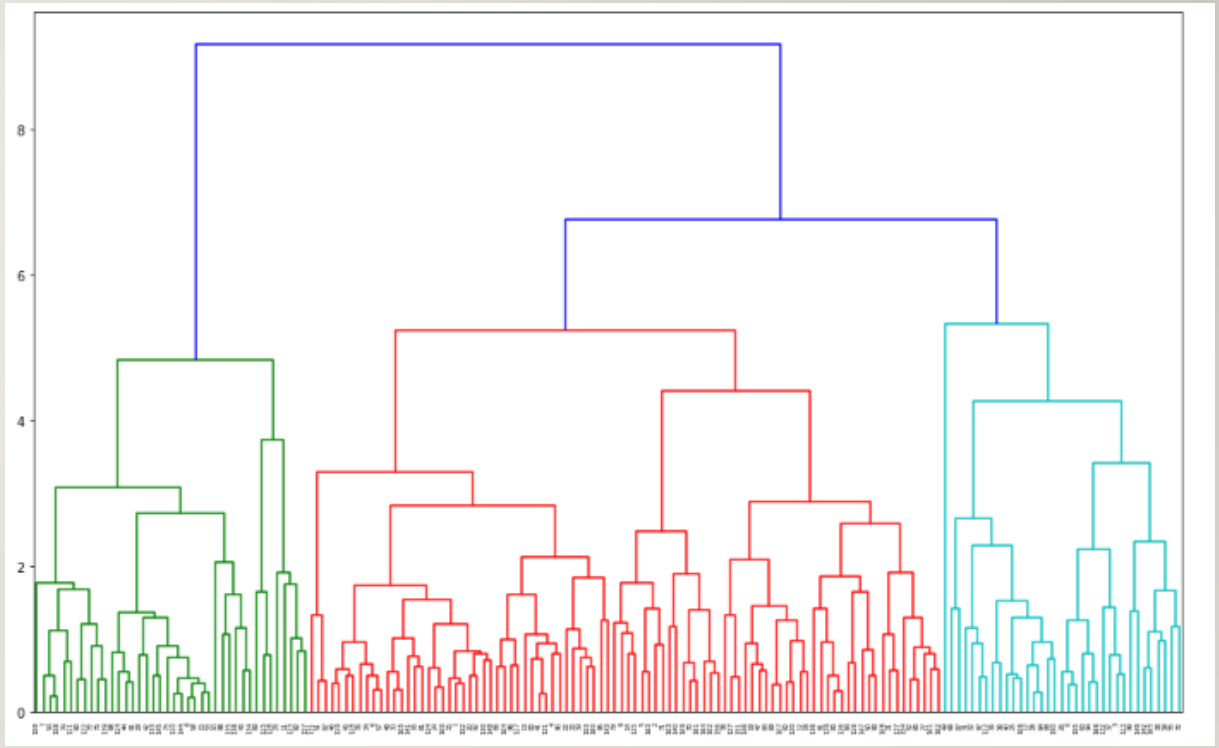


It shows higher the income and gdpp of a country, lower is the child mortality rate due to better health care, less inflation.

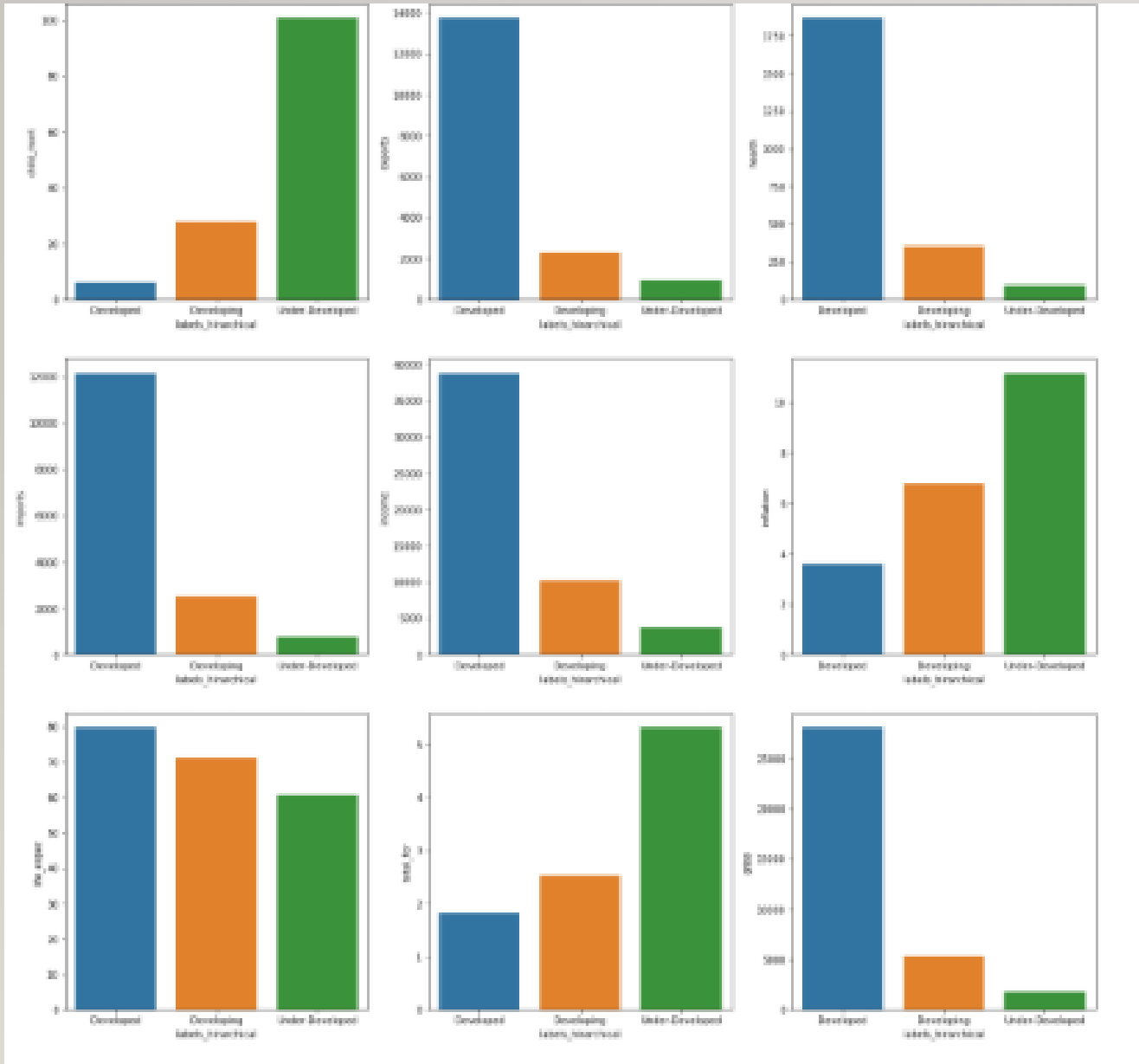
Hierarchical Clustering Single Linkage



Hierarchical Clustering Complete Linkage

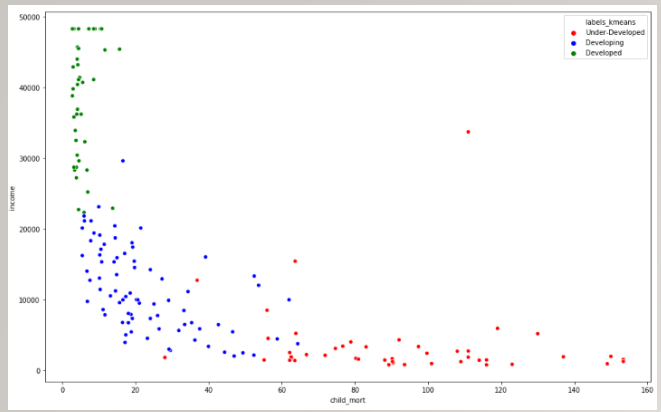


We can see how single and complete linkage stack against each other in Hierarchical clustering. Clustering points are taken as 3 through cut tree algorithm.

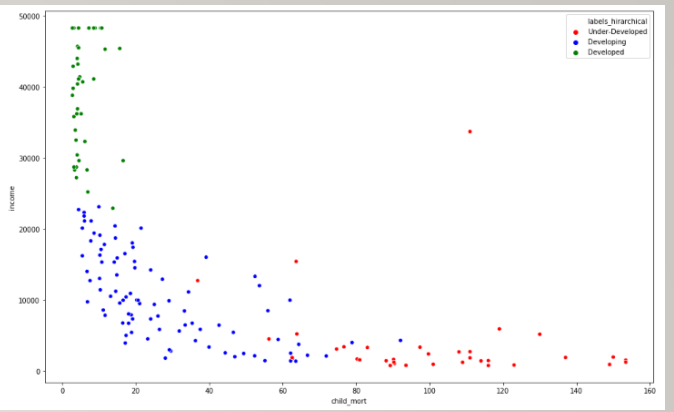


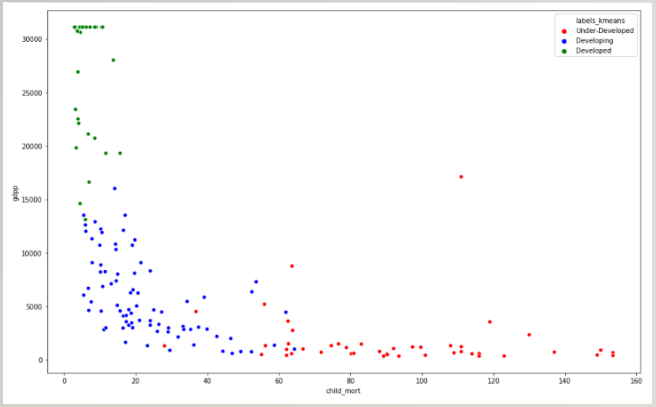
We can see how all the variables are stacked against the labels showing with developed, developing and underdeveloped countries, respectively.

We can see how GDP plays an important role in the development of the country. Higher the income, higher the imports and exports, lesser the inflation, better is the health of people, higher the life expectancy.

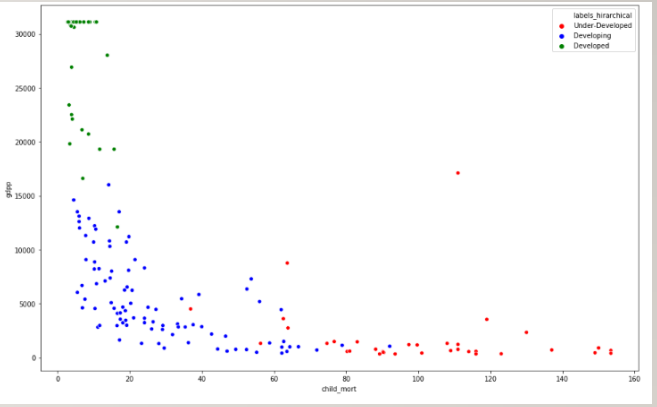


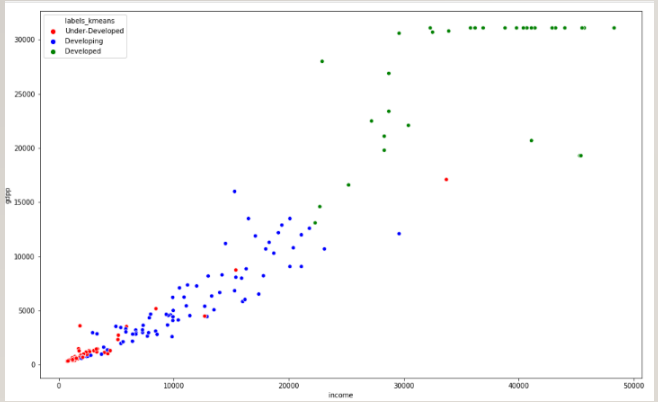
Visualizing both K-Means and Hierarchical Clustering at left and right respectively, the first graph on both sides show how income and child_mort are inversely correlated with each other. Higher the income, lower is the child mortality.



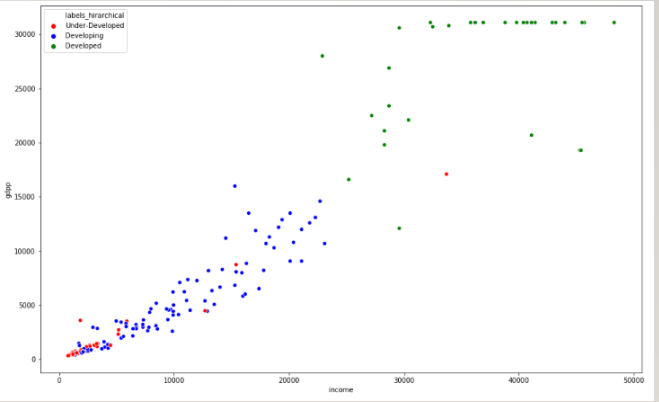


In the second graph we can see that how gdpp and child_mort are inversely related to each other and it's the same outcome, higher the gdpp, lower is the child mortality.





In the third graph we can see, how directly proportional are income and gdpp. Higher the income, higher the gdpp.



K-Means Clustering

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	labels_kmeans
26	Burundi	93.6	22.243716	26.7960	104.90964	764.00	12.30	57.82	6.2600	331.62	0
88	Liberia	89.3	62.457000	38.5860	302.80200	742.24	5.47	60.80	5.0200	331.62	0
37	Congo, Dem. Rep.	116.0	137.274000	26.4194	165.66400	742.24	20.80	57.82	6.5400	334.00	0
112	Niger	123.0	77.256000	17.9568	170.86800	814.00	2.55	58.80	6.5636	348.00	0
132	Sierra Leone	153.4	67.032000	52.2690	137.65500	1220.00	17.20	57.82	5.2000	399.00	0

Hierarchical Clustering

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	labels_kmeans	labels_hirarchical
26	Burundi	93.6	22.243716	26.7960	104.90964	764.00	12.30	57.82	6.2600	331.62	Under-Developed	0
88	Liberia	89.3	62.457000	38.5860	302.80200	742.24	5.47	60.80	5.0200	331.62	Under-Developed	0
37	Congo, Dem. Rep.	116.0	137.274000	26.4194	165.66400	742.24	20.80	57.82	6.5400	334.00	Under-Developed	0
112	Niger	123.0	77.256000	17.9568	170.86800	814.00	2.55	58.80	6.5636	348.00	Under-Developed	0
132	Sierra Leone	153.4	67.032000	52.2690	137.65500	1220.00	17.20	57.82	5.2000	399.00	Under-Developed	0

We can see that from both the k-means and hierarchical clustering, the top 5 countries which needs support from NGO.

Final Thoughts:

The results vary accordingly depending on how we treat outliers. Capping has been done in such a way that we do not lose any important limits be it upper limit in child mortality column or lower limits in income and GDP.

Ultimately, we get the top 5 countries which are in dire need of help, because of its low income, low GDP, low health care, high child mortality rate and low life expectancy.

Top 5 countries which are in dire need of help from NGO are:

1. Burundi: with 'child_mort' = 93.6, 'income' = 764, 'gdpp' = 331.62
2. Liberia: with 'child_mort' = 89.3, 'income' = 742.24, 'gdpp' = 331.62
3. Congo, Dem. Rep.: with 'child_mort' = 116, 'income' = 742.24, 'gdpp' = 334
4. Niger: with 'child_mort' = 116, 'income' = 814, 'gdpp' = 348
5. Sierra Leone: with 'child_mort' = 116, 'income' = 1220, 'gdpp' = 399