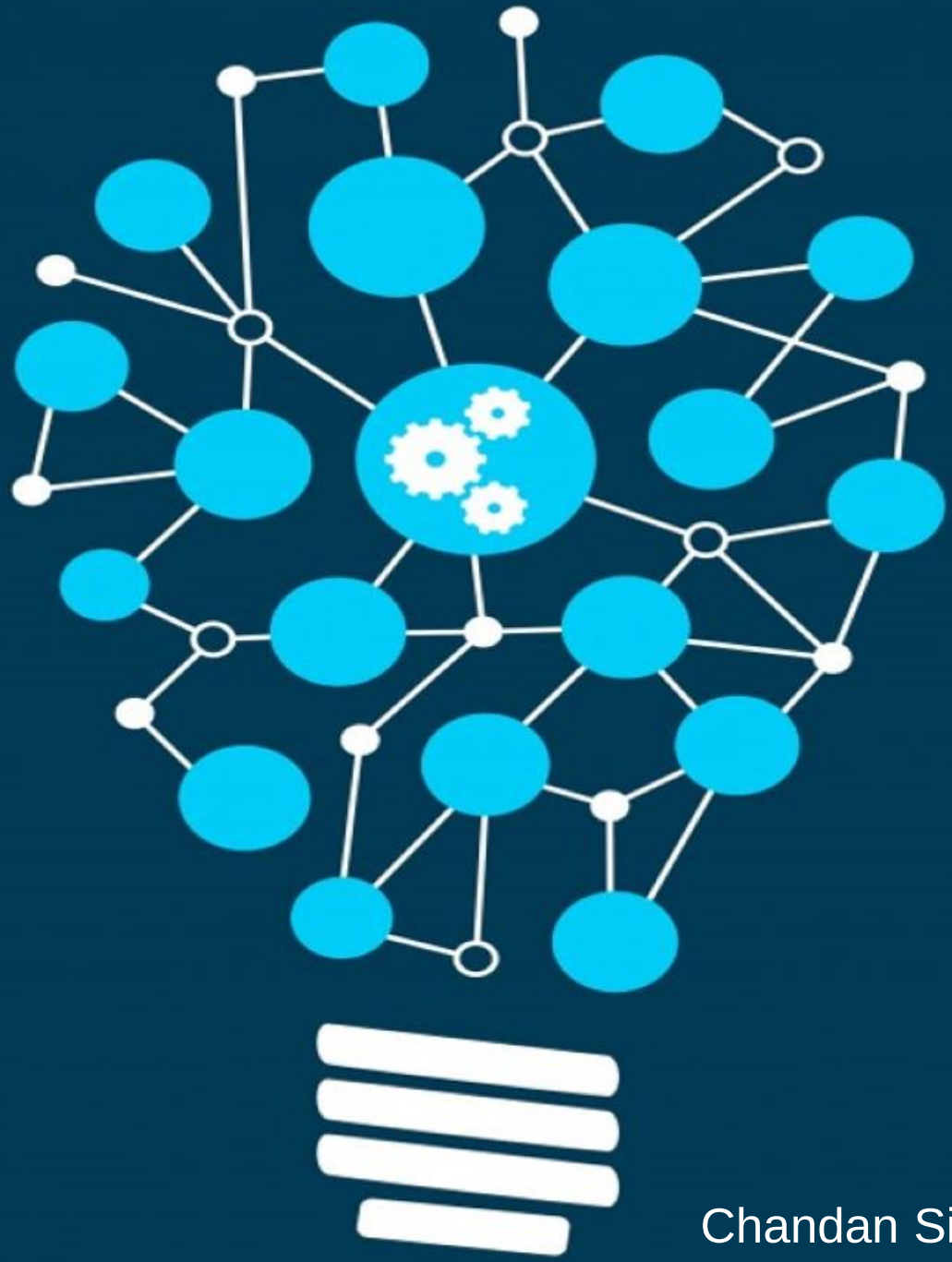
Clustering And PCA Assignment

# MACHINE LEARNING



Chandan Singh

# 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. It runs a lot of operational projects from time to time along with advocacy drives to raise awareness as well as for funding purposes.*

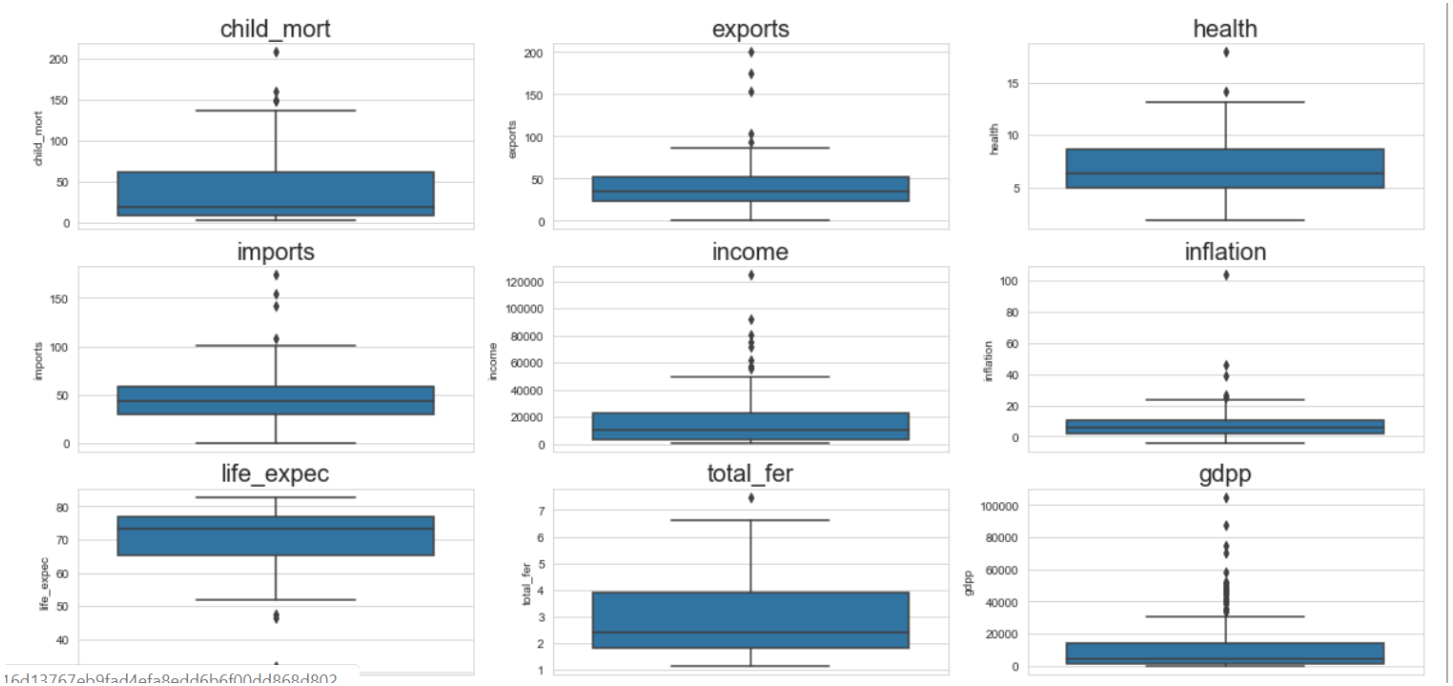
*After the recent funding programmes, they have been able to raise around $10 million. Now the CEO of the NGO needs to decide how to use this money strategically and effectively. The significant issues that come while making this decision are mostly related to choosing the countries that are in the direst need of aid.*

# Analysis Approach

The steps are broadly:

1. Read and understand the dataset provided
2. Clean the data
3. Outliers analysis and treatment
4. Scale the features
5. Perform PCA(Principal Component Analysis) on the data
6. Clustering
    i. K-Means Clustering
    ii. Hierarchical Clustering
7. Analyse and Visualise the principal components
8. Analyse and Visualise the original variables
9. Final Analysis and recommendations

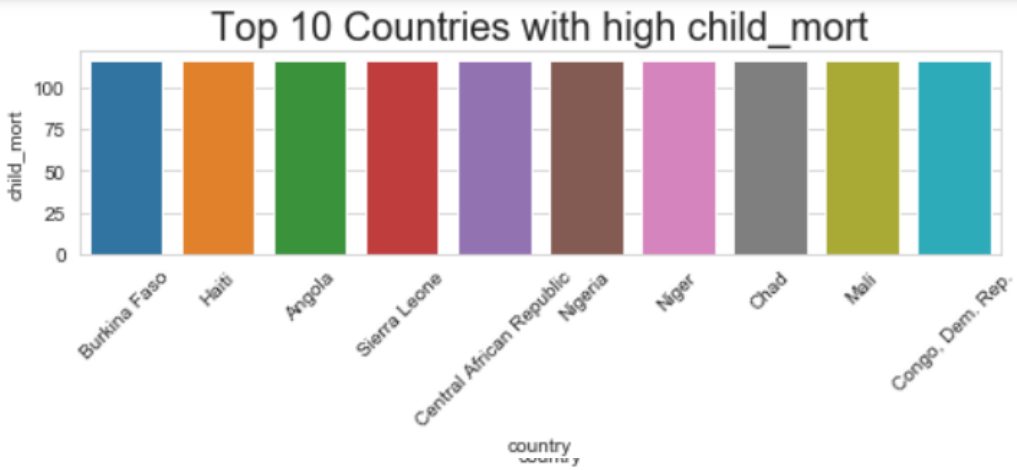
# Outliers Analysis

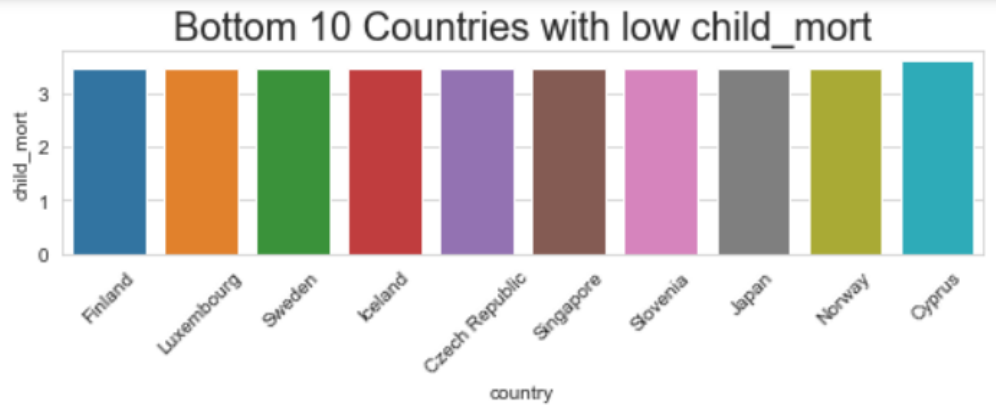


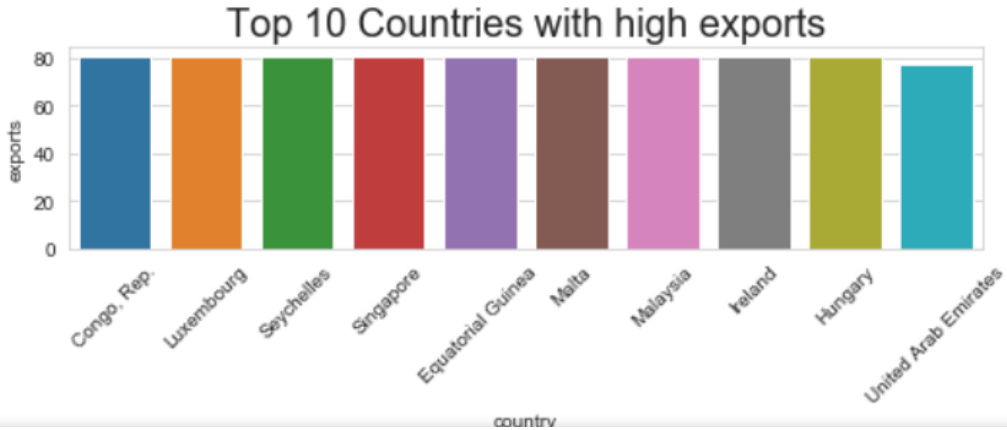

It can be clearly seen that there are outliers present in the columns, especially **gdpp, child_mort, inflation, Income** etc. Since, we need to perform the analysis, based on the socio and economy factors, hence, it is not recommended to remove the countries, that are outliers.

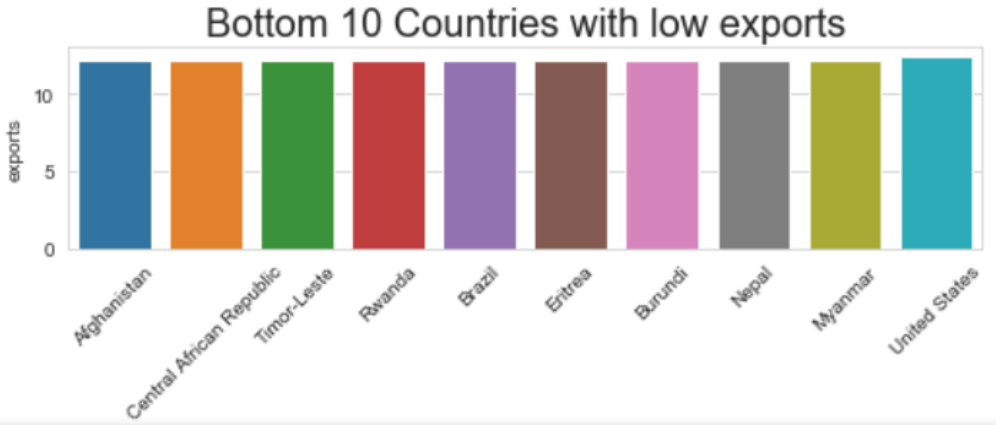
A good practice is to cap them to avoid any impact.

# Socio-Economic Factor Analysis

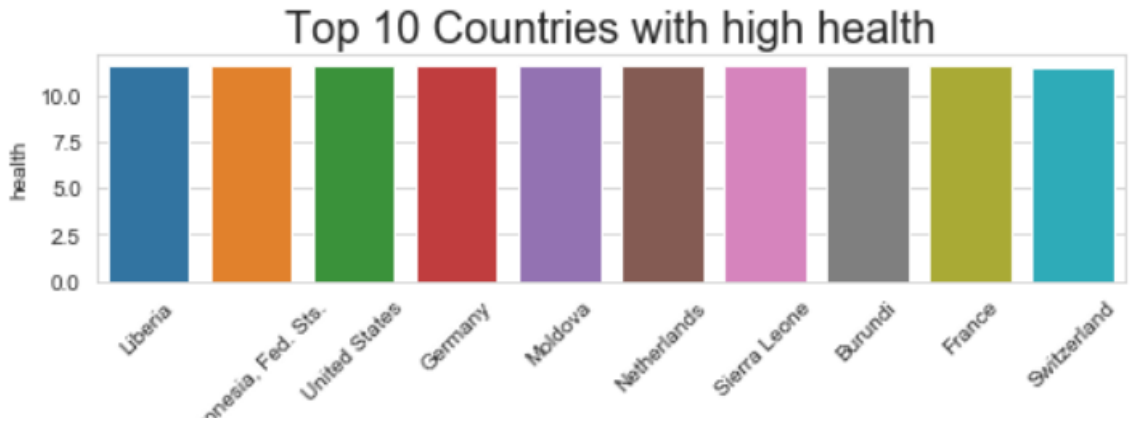


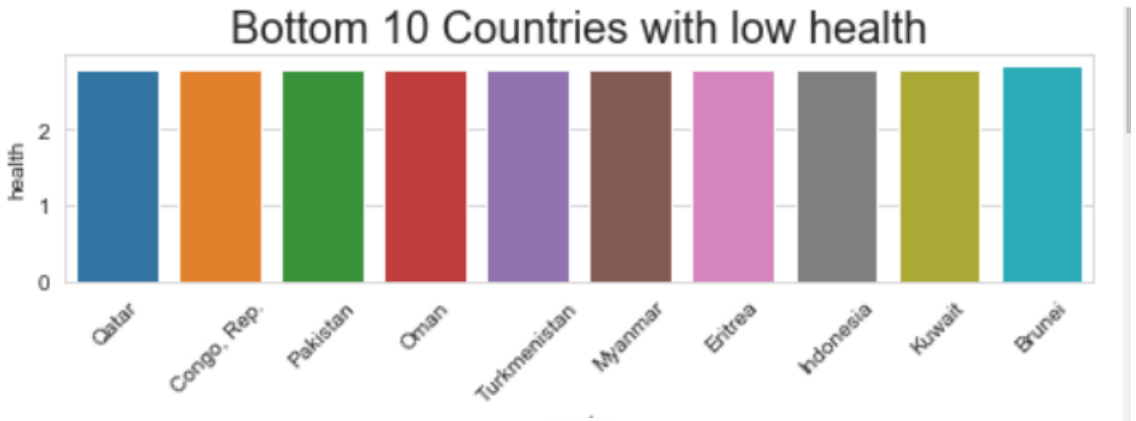



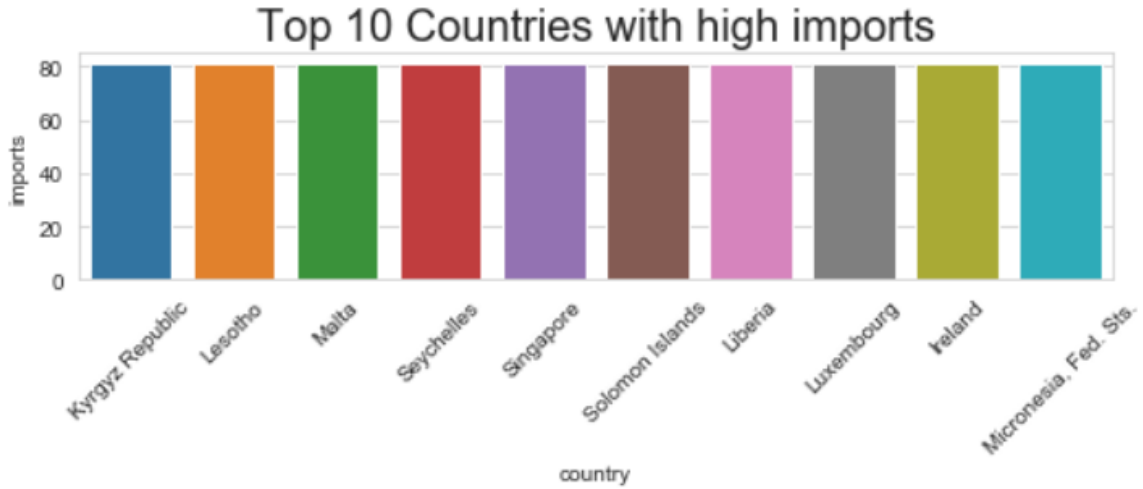



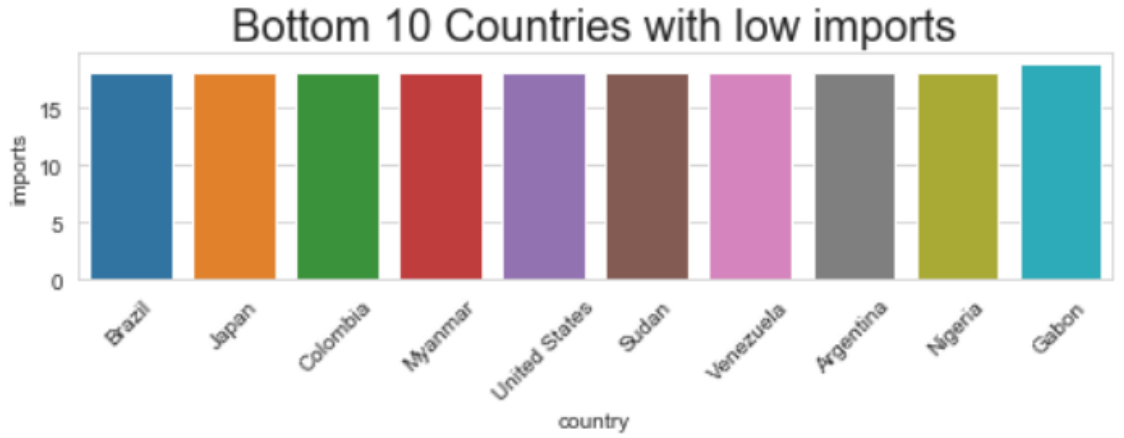



# Continued..

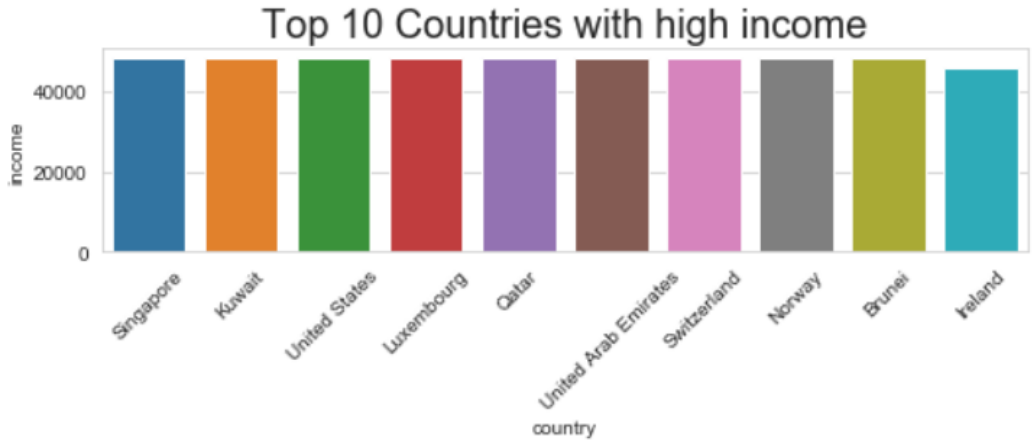

Top 10 Countries with high health
health
10.0
7.5
5.0
2.5
0.0
Liberia
United States
Germany
Moldova
Netherlands
Sierra Leone
Burundi
France
Switzerland

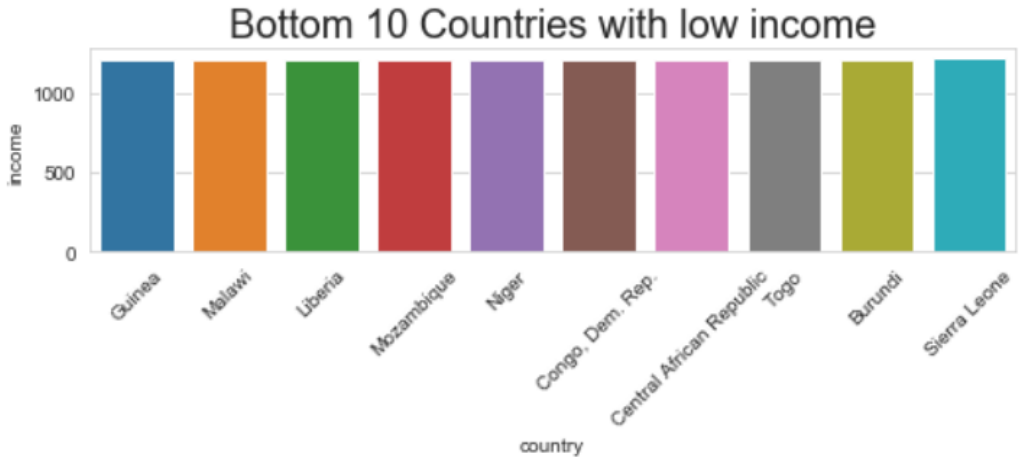

Bottom 10 Countries with low health
health
2
1
0
Qatar
Congo, Rep.
Pakistan
Oman
Turkmenistan
Myanmar
Eritrea
Indonesia
Kuwait
Brunei

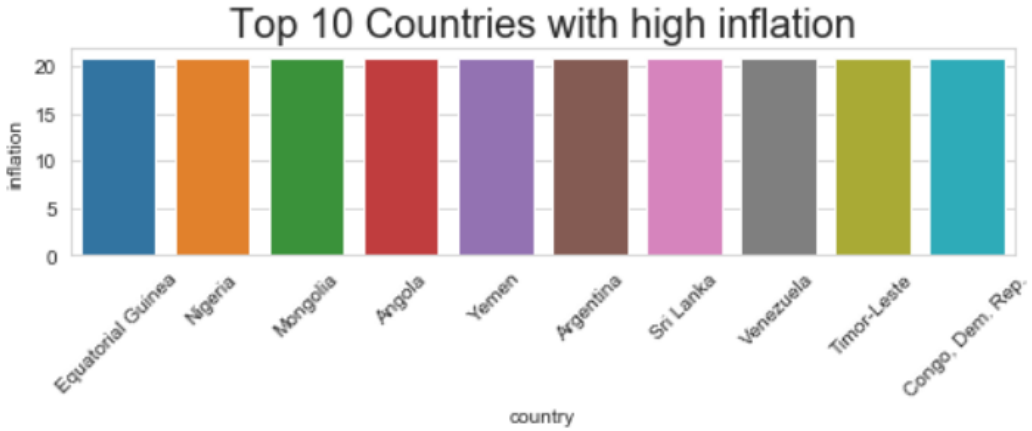

Top 10 Countries with high imports
imports
80
60
40
20
0
Kyrgyz Republic
Lesotho
Malta
Seychelles
Singapore
Solomon Islands
Liberia
Luxembourg
Ireland
Micronesia, Fed. Sts.
country

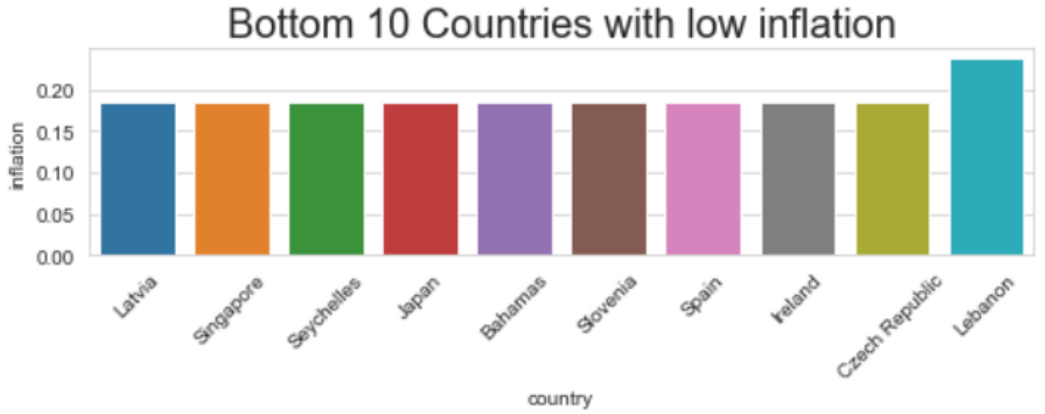

Bottom 10 Countries with low imports
imports
15
10
5
0
Brazil
Japan
Colombia
Myanmar
United States
Sudan
Venezuela
Argentina
Nigeria
Gabon
country

# Continued..

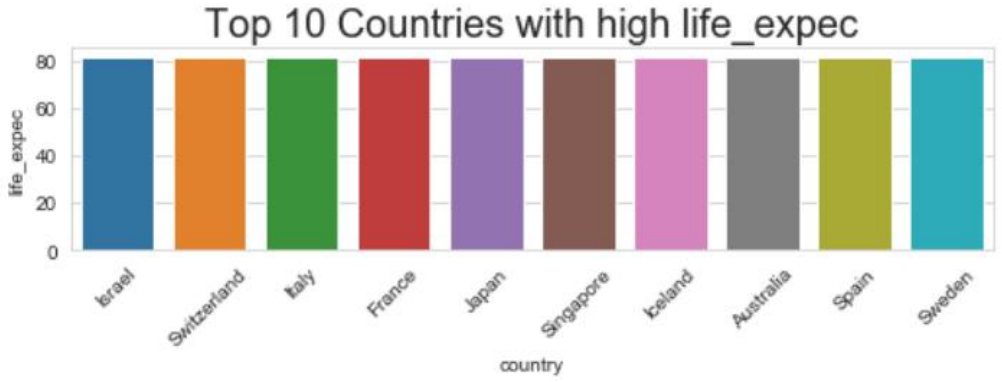

Top 10 Countries with high income
income
40000
20000
0
Singapore
Kuwait
United States
Luxembourg
Qatar
United Arab Emirates
Switzerland
Norway
Brunei
Ireland
country

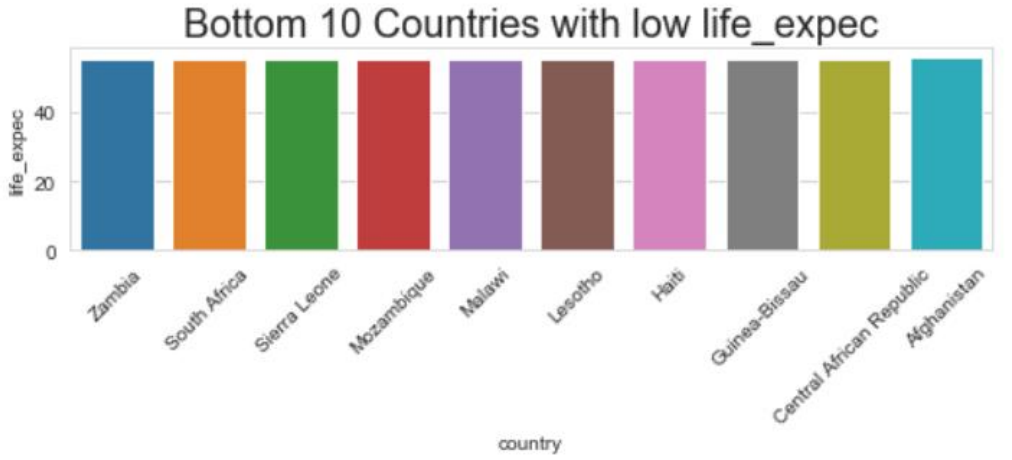

Bottom 10 Countries with low income
income
1000
500
0
Guinea
Malawi
Liberia
Mozambique
Niger
Congo, Dem. Rep.
Central African Republic
Togo
Burundi
Sierra Leone
country

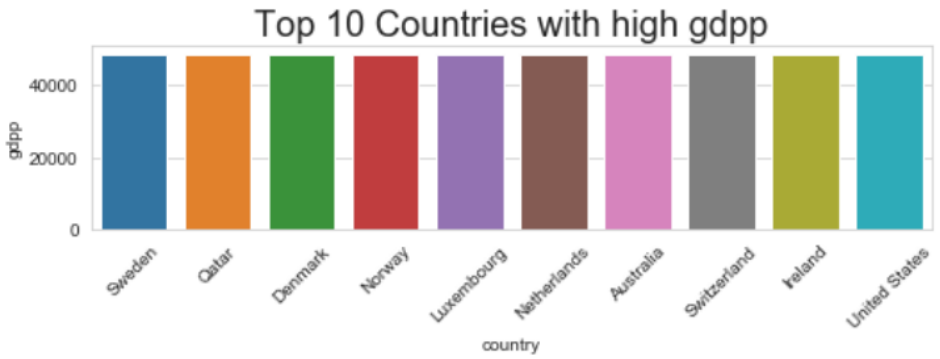

Top 10 Countries with high inflation
inflation
20
15
10
5
0
Equatorial Guinea
Nigeria
Mongolia
Angola
Yemen
Argentina
Sri Lanka
Venezuela
Timor-Leste
Congo, Dem. Rep.
country

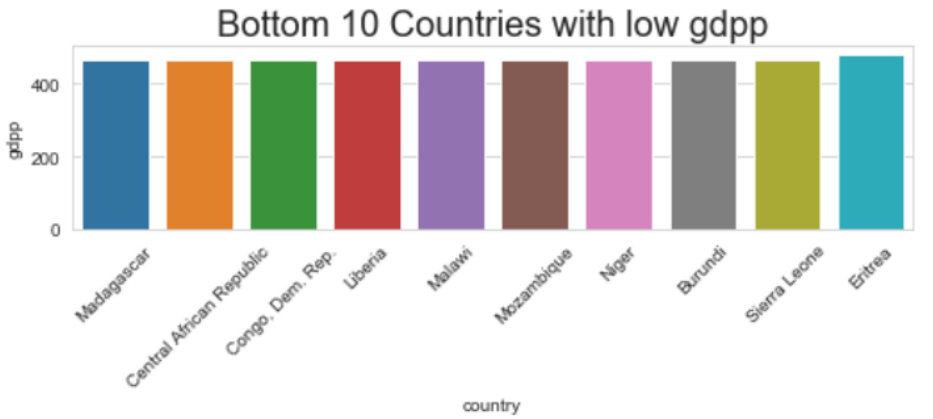

Bottom 10 Countries with low inflation
inflation
0.20
0.15
0.10
0.05
0.00
Latvia
Singapore
Seychelles
Japan
Bahamas
Slovenia
Spain
Ireland
Czech Republic
Lebanon
country

# Continued..

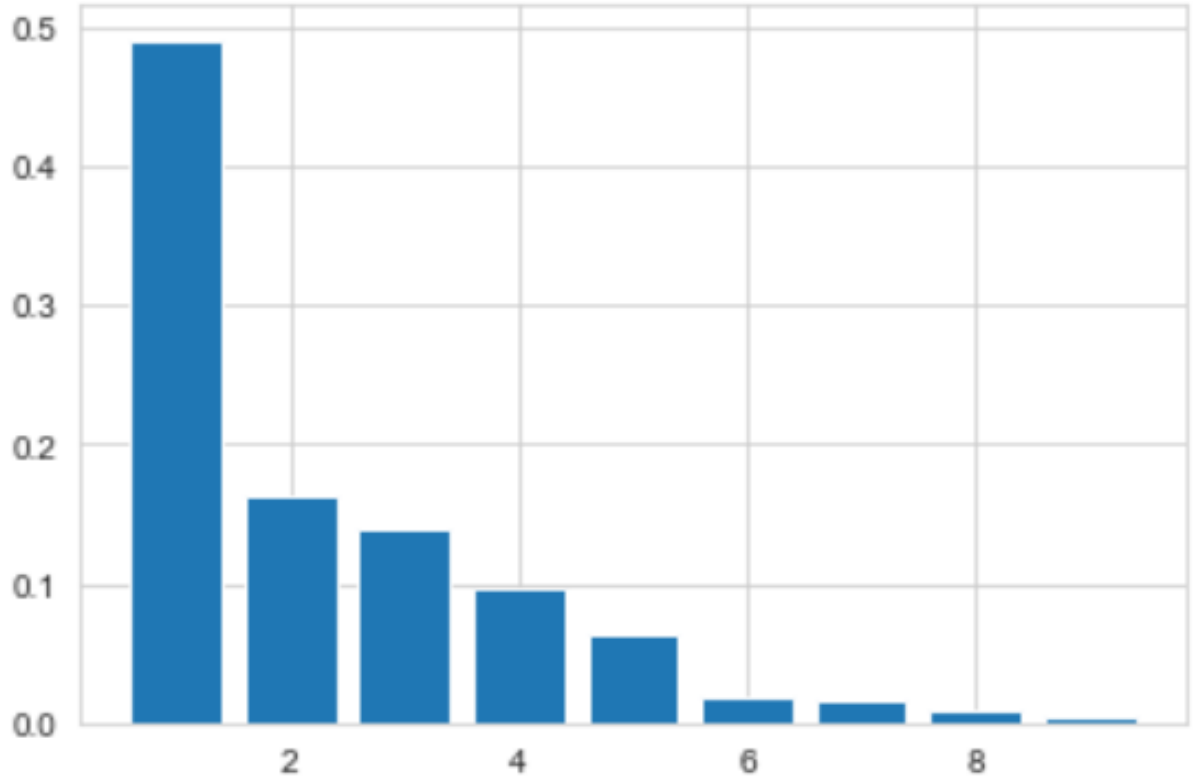

Top 10 Countries with high life_expec
life_expec
0
20
40
60
80
Israel
Switzerland
Italy
France
Japan
Singapore
Iceland
Australia
Spain
Sweden
country


Bottom 10 Countries with low life_expec
life_expec
0
20
40
Zambia
South Africa
Sierra Leone
Mozambique
Malawi
Lesotho
Haiti
Guinea-Bissau
Central African Republic
Afghanistan
country


Top 10 Countries with high gdpp
gdpp
0
20000
40000
Sweden
Qatar
Denmark
Norway
Luxembourg
Netherlands
Australia
Switzerland
Ireland
United States
country

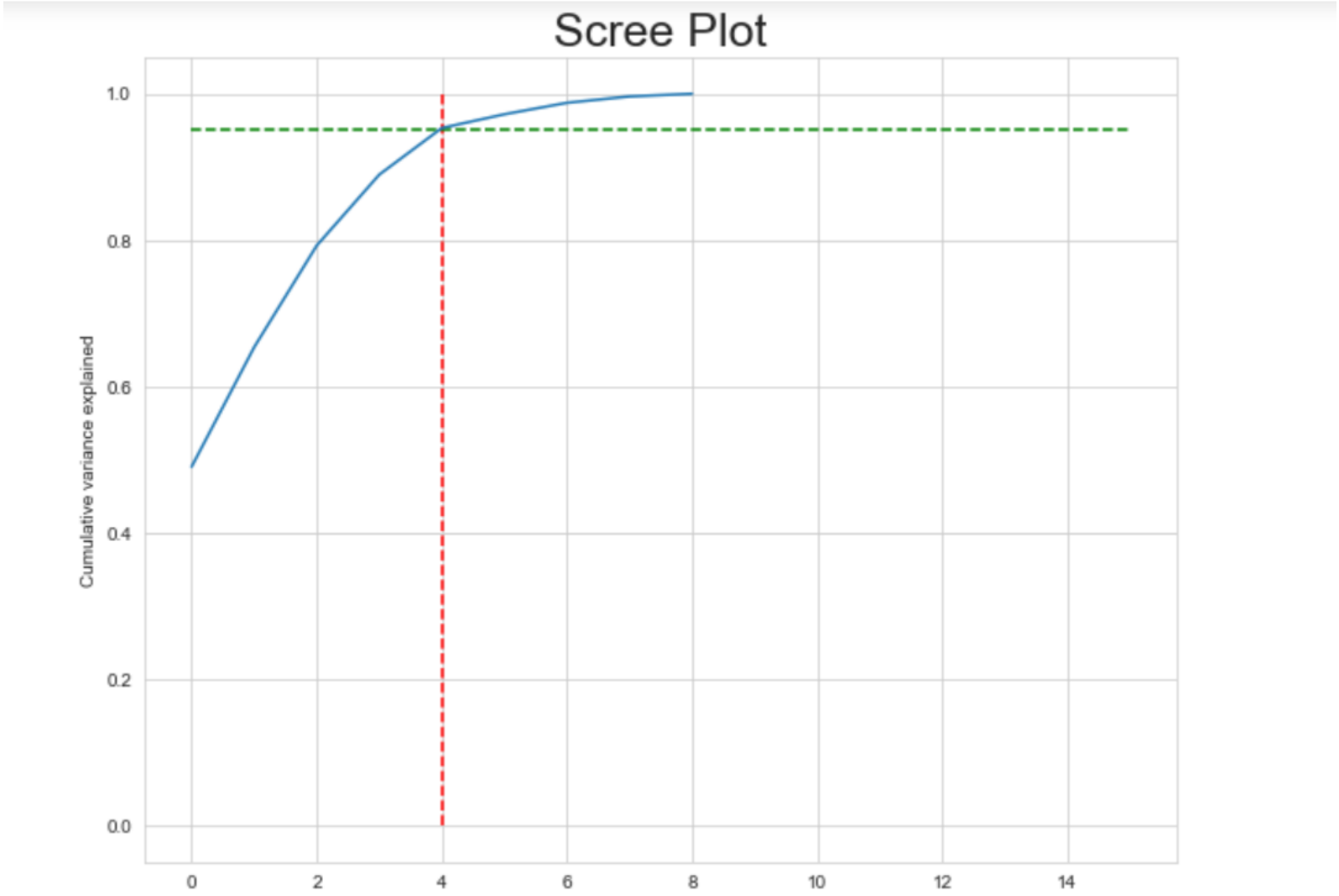

Bottom 10 Countries with low gdpp
gdpp
0
200
400
Madagascar
Central African Republic
Congo, Dem. Rep.
Liberia
Malawi
Mozambique
Niger
Burundi
Sierra Leone
Eritrea
country

# PCA Analysis



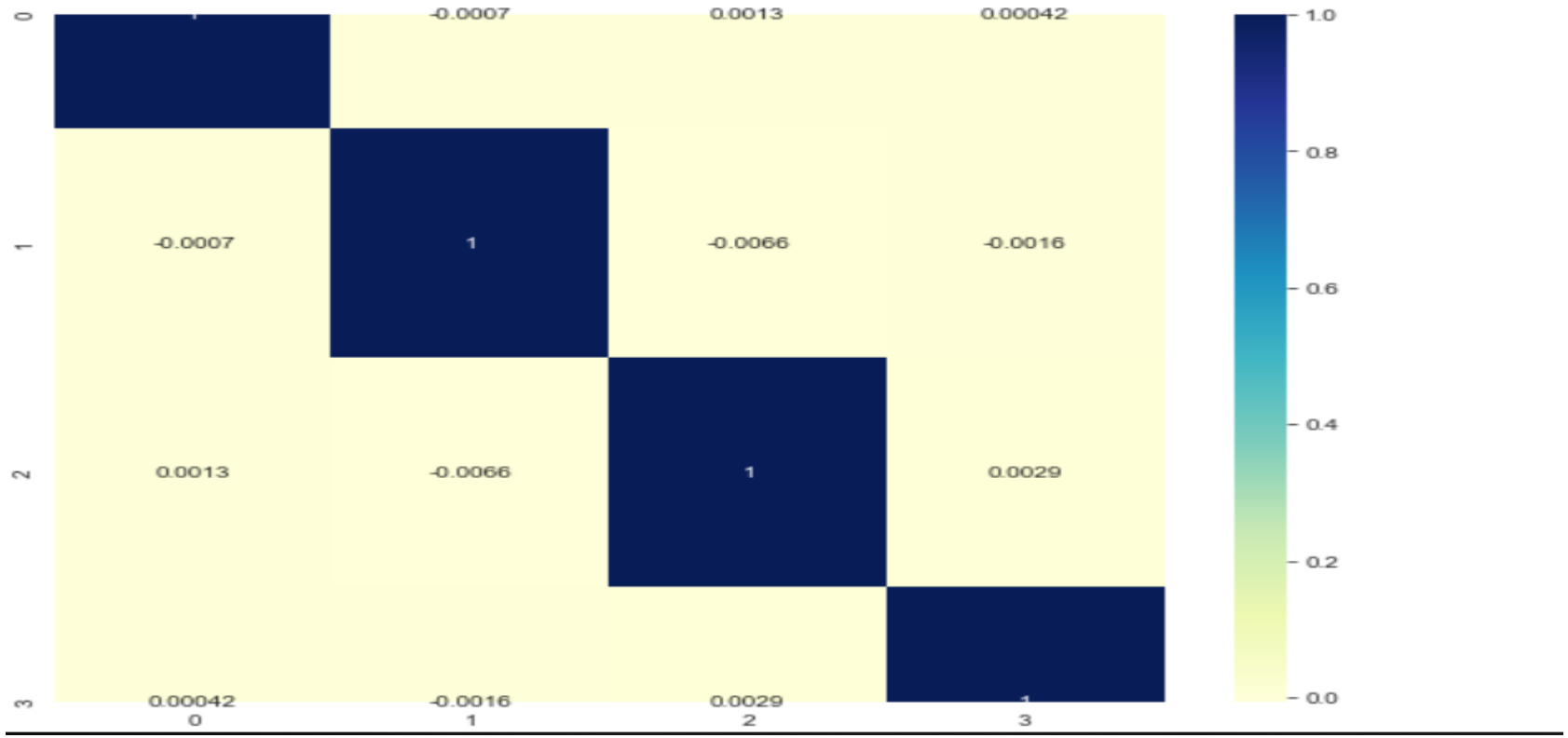

- Out of 9 variables, 95% variance is explained by 4 variables.
- These 4 variables capture the maximum information.
- These variables are also called Principal Components

# Continued..


Scree Plot
Cumulative variance explained
1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
8
10
12
14

# Continued..

## Principal Components Correlation




- It is clearly seen that all the Principal Components are uncorrelated
- All the Principal components have almost zero correlation

# Clustering : K-Means

## Silhouette Analysis

$$\text{silhouette score} = \frac{p - q}{max(p, q)}$$

$p$ 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.

Silhouette score for 4 clusters is **0.412**

# Continued..

## Hopkins Statistics:

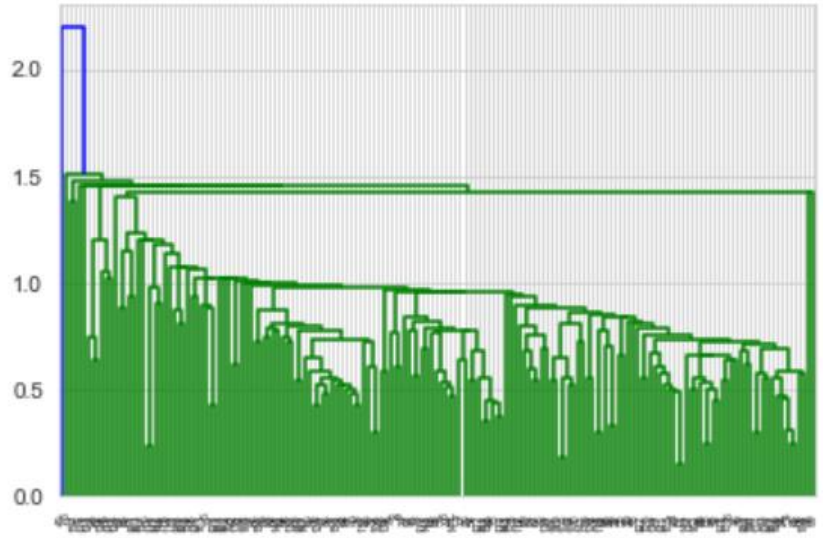
The Hopkins statistic, is a statistic which gives a value which indicates the cluster tendency, in other words: how well the data can be clustered.

- If the value is between {0.01, ...,0.3}, the data is regularly spaced.
- If the value is around 0.5, it is random.
- If the value is between {0.7, ..., 0.99}, it has a high tendency to cluster.

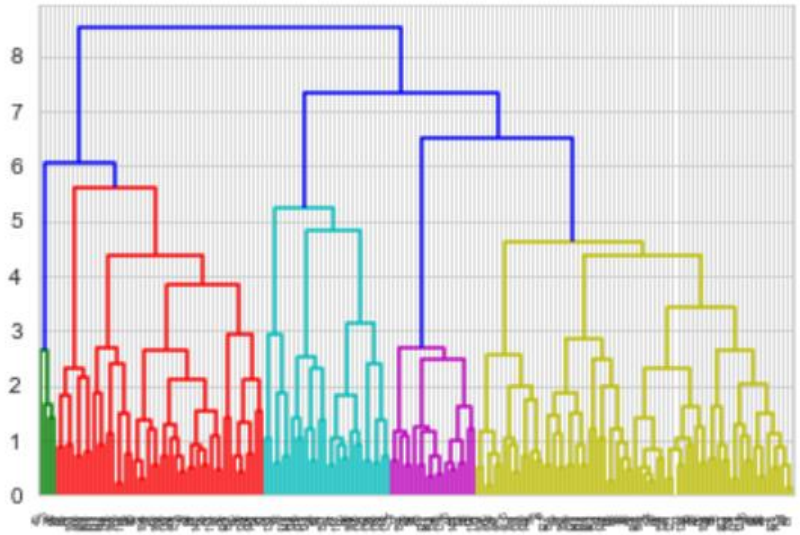
Hopkins Score for the dataset is **0.839**

# Clustering : Hierarchical

Single Linkage



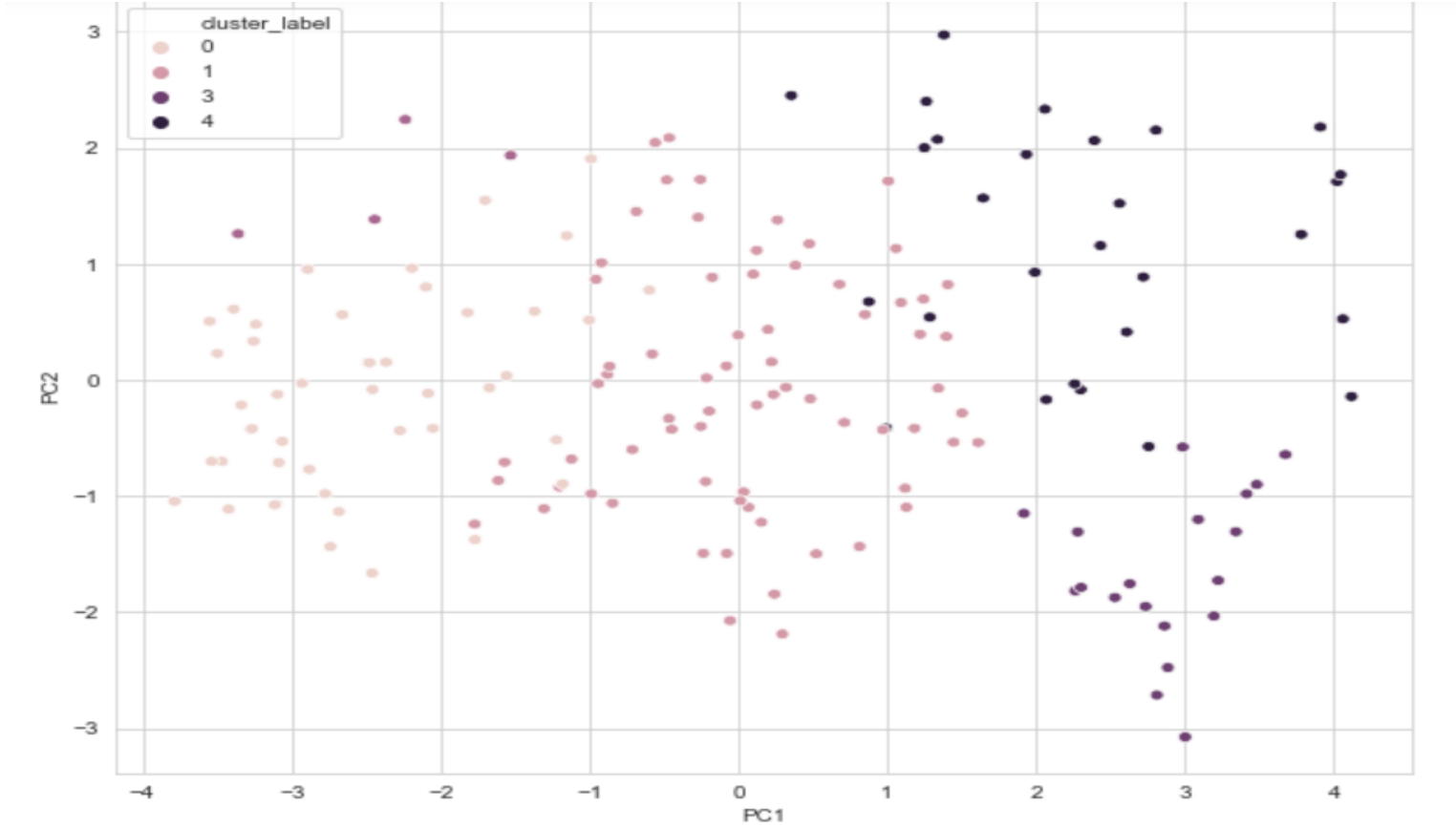

Complete Linkage




- Dendrogram cutting edges across 5 clusters
- Dividing countries into 5 groups/clusters is recommended

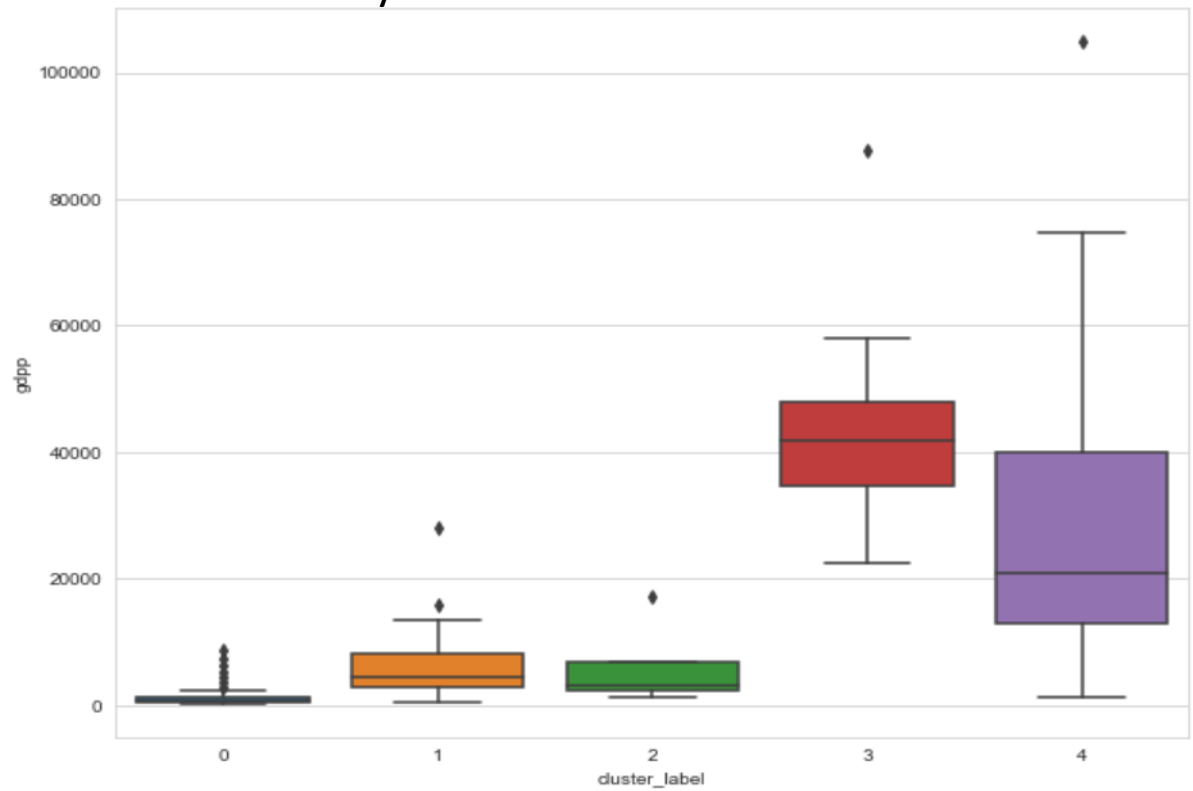
# Clustering : Visualisation

Scatter Plot of first 2 Principal Components and Clusters



# Clustering : GDPP Analysis

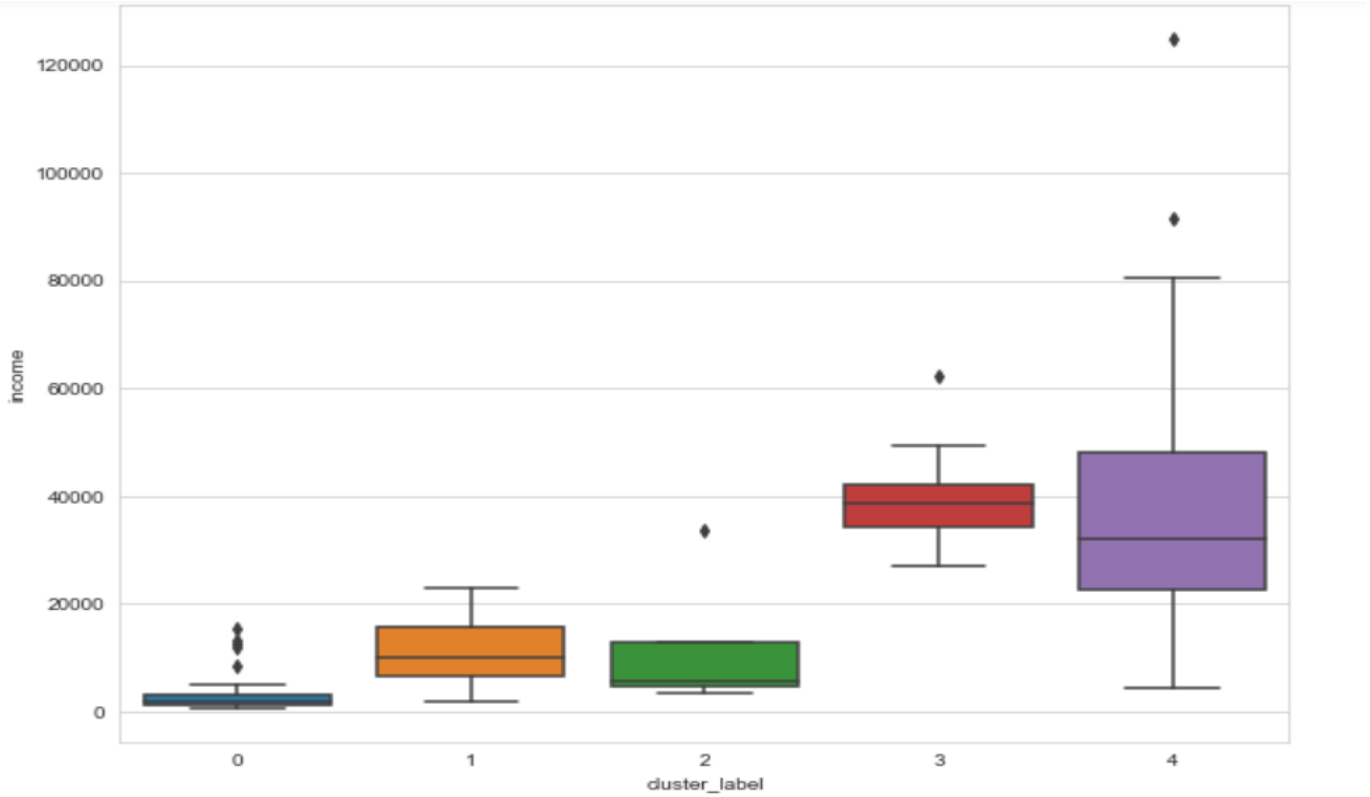
Cluster Analysis of Variable GDPP




- Countries that belong to Cluster_Label '0' have the least GDPP.
- These countries can be considered as backward countries.
- Countries that belong to Clusger_Label '4' have the highest GDPP.

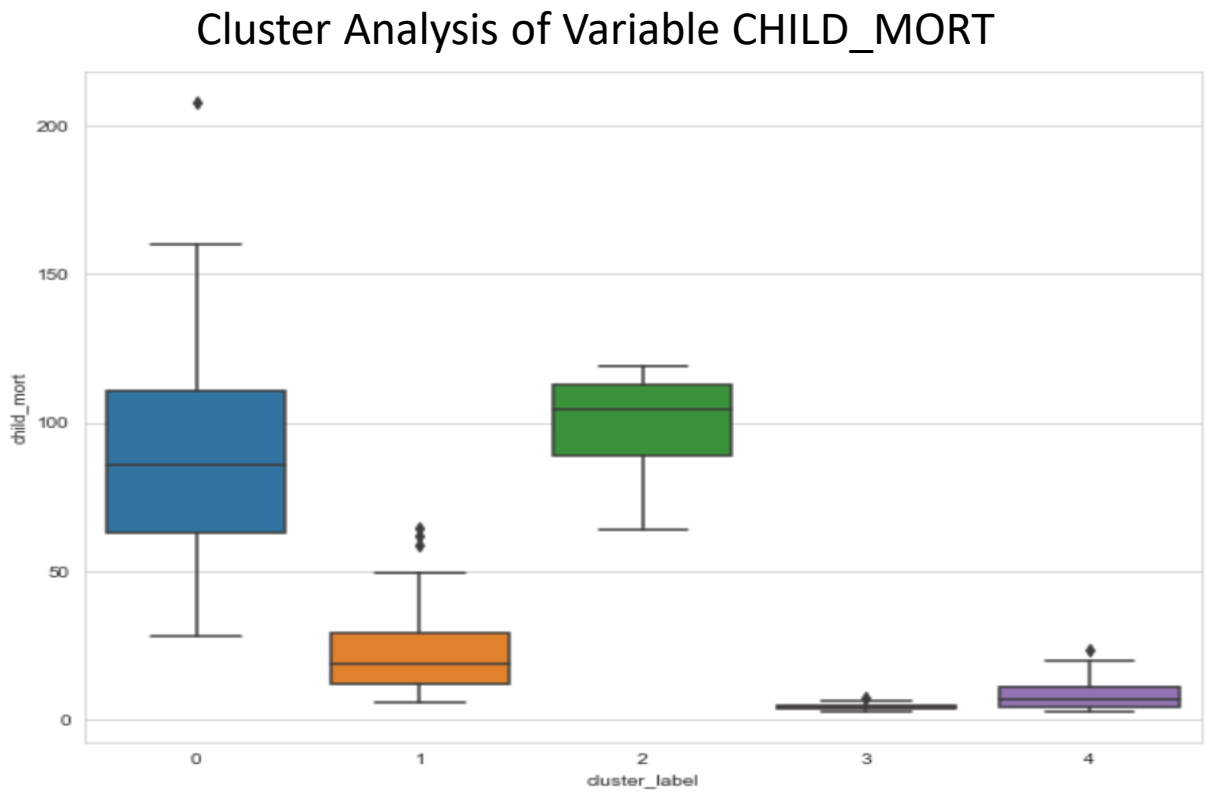
# Clustering : Income Analysis

Cluster Analysis of Variable INCOME




- Countries that belong to Cluster_Label '0' have the least net income. Thus, these countries can be considered as backward.
- Countries that belong to Cluster_Label '4' have the highest net income and are very well developed countries.

# Clustering : CHILD_MORT Analysis




- It is evident that the countries belong to Cluster_Label '0' have the highest child mortality rate.
- Countries in Cluster_Label '3' have the least child mortality rate.

# Recommendations

As per the analysis and evidence from the graphs, the countries that belong to Cluster_Label '0' have:

- Low GDPP
- Low Net Income
- High Child Mortality

Therefore, these countries are backward countries and are in dire need of aid. NGOs should focus more on these countries:

The list of the countries :

| | | | |
|---|---|---|---|
| Afghanistan | Cote d'Ivoire | Liberia | Senegal |
| Benin | Eritrea | Madagascar | Sierra Leone |
| Botswana | Gambia | Malawi | Solomon Islands |
| Burkina Faso | Guinea-Bissau | Mali | South Africa |
| Burundi | Haiti | Micronesia, Fed. Sts. | Tanzania |
| Cameroon | Iraq | Mozambique | Timor-Leste |
| Central African Republic | Kenya | Namibia | Togo |
| Chad | Kiribati | Niger | Uganda |
| Comoros | Lao | Pakistan | Zambia |
| Congo, Dem. Rep. | Lesotho | Rwanda | |