Clustering Assignment

by Devendra Kumar

Objective and Problem Statement

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

Problem Statement: We as a Data Analyst, have to come up with the top 5 countries which are in serious need of the AID on the basis of their socio-economy and health factors.

Solution Methodology

Data Extraction
- Importing libraries
- Loading data
- Observing shape of data.

EDA or Data Visualization
- Correlation plot
- Pair plot
- Univariate and Bivariate analysis using distplot and boxplot

Data Preparation
- Identifying missing values.
- Identifying continuous and categorical columns.
- Identifying outliers.

Outliers Treatment
- Capping Outliers

Feature Scaling
- Scaling data using StandardScaler

Hopkin's Statistic Test
- Cluster tendency

Solution Methodology Contt..

K-Means Clustering

- Identifying the k using silhouette and elbow-SSD curve.
- Labeling data using k=3.
- Assigning labels to dataframe.
- Observing type of clusters formed with k=3.
- Profiling clusters on basis of gdpp, child_mort and income rate.

Hierarchical Clustering

- Identifying the k using single and complete linkage
- Labeling data using k=3 from Dendrogram
- Assigning labels to dataframe.
- Observing type of clusters formed with k=3.
- Profiling clusters on basis of gdpp, child_mort and income rate.

Choosing the appropriate Clustering Technique.

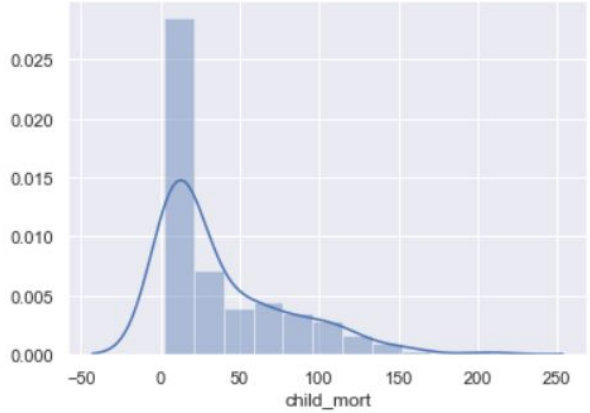
Decision Making

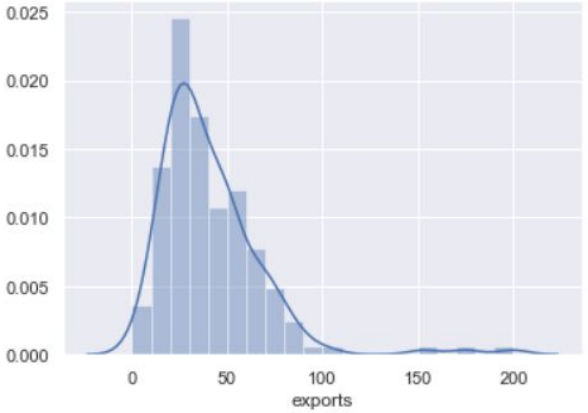
- Identifying top 5 countries which are in need of AID
- Visual analysis of top 5 countries

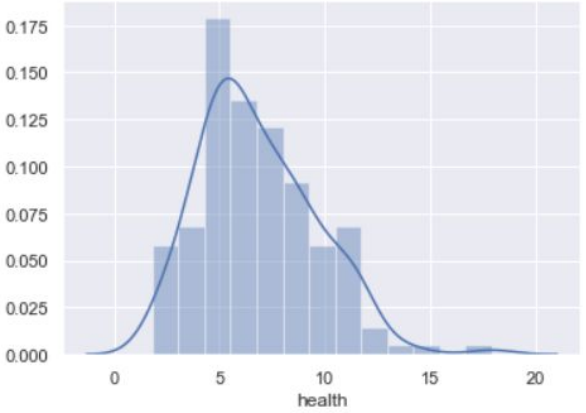
Correlation in Data

- From correlation plot we can see that few features highly correlated with each other for example gdpp and imports, child_mort and total_fer, income and exports.
- As all other features are highly correlated thats why I guess we are choosing the only 3 features for profiling the clusters i.e. gdpp, child_mort and income rate.

Continuous variables analysis

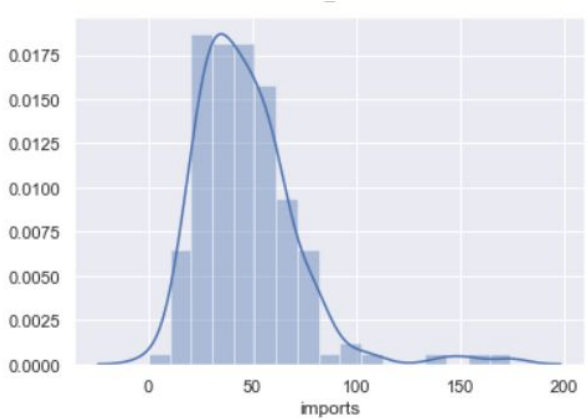


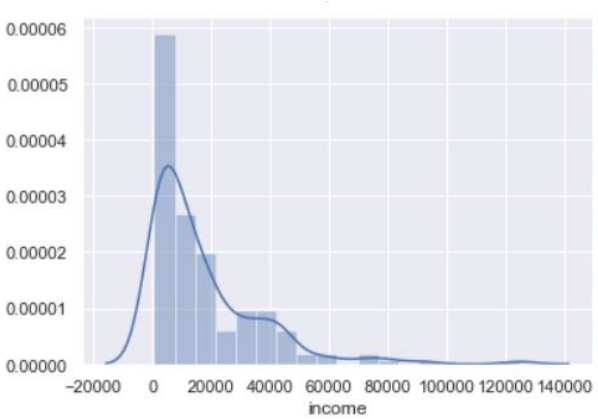


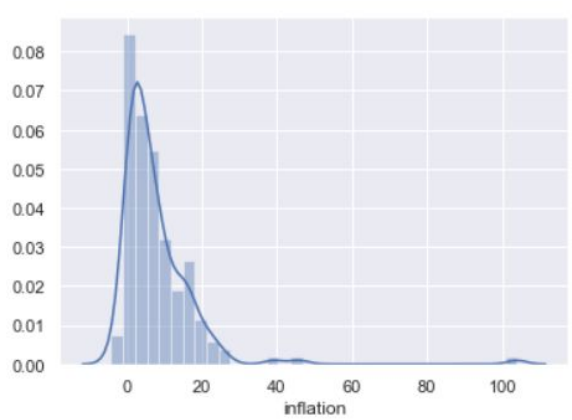


- **child_mort:** Here we see, that the graph on its peak on low values of **child_mort**, Hence, we can interpret that the maximum countries has **low child_mort.**
- **exports** and **health** are finely distributed among all countries as we see the graph is almost normal distributed. Hence, we can say that these variables will not help us to analysing the countries are in need of AID so, not taking these values for profiling the clusters.

Continuous variables analysis contt...

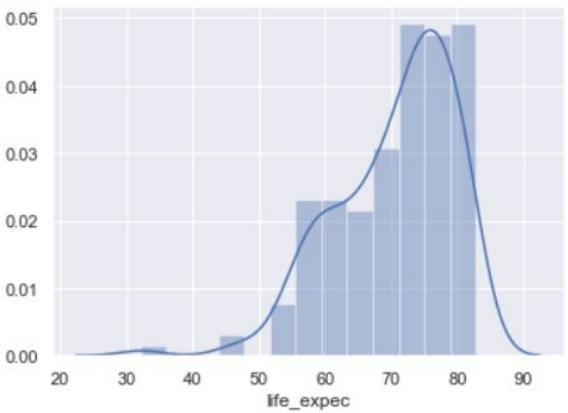


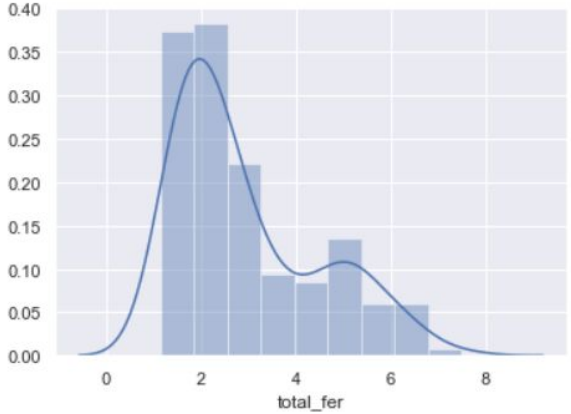


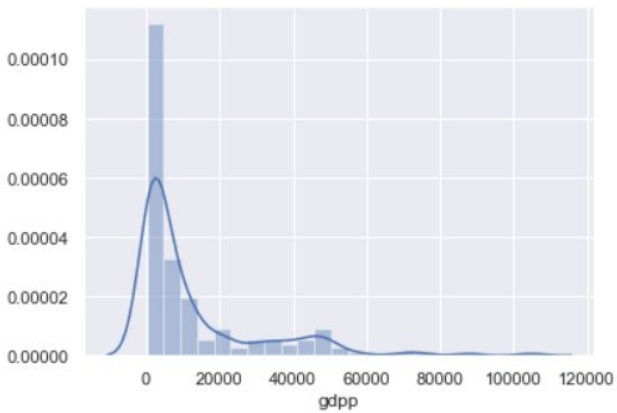


- **Imports** are almost normally distribute among clusters.
- **Income** looks very interesting to taking part in cluster profiling as it is having ups and downs in between values. Highly countries are having average income **~10k to ~20k.**
- **Inflation** is avg of ~10 units among countries and there is nearly zero inflation also for some of the countries. This variable could be a good profiling variable.

Continuous variables analysis contt...

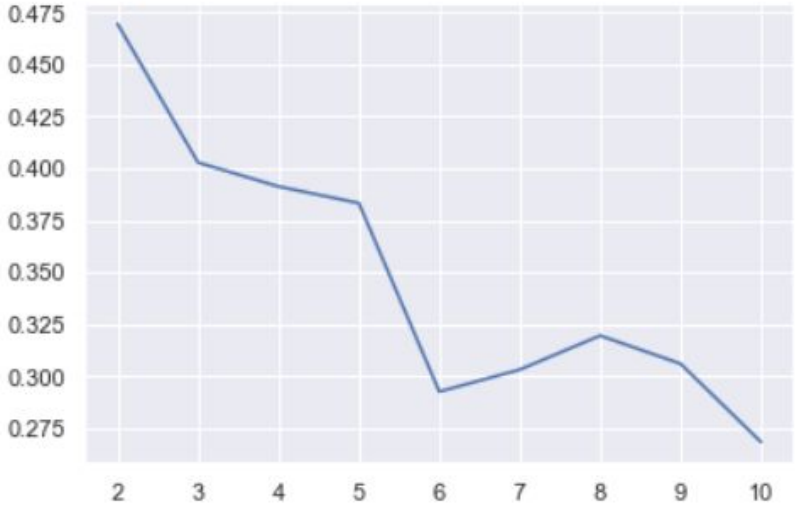




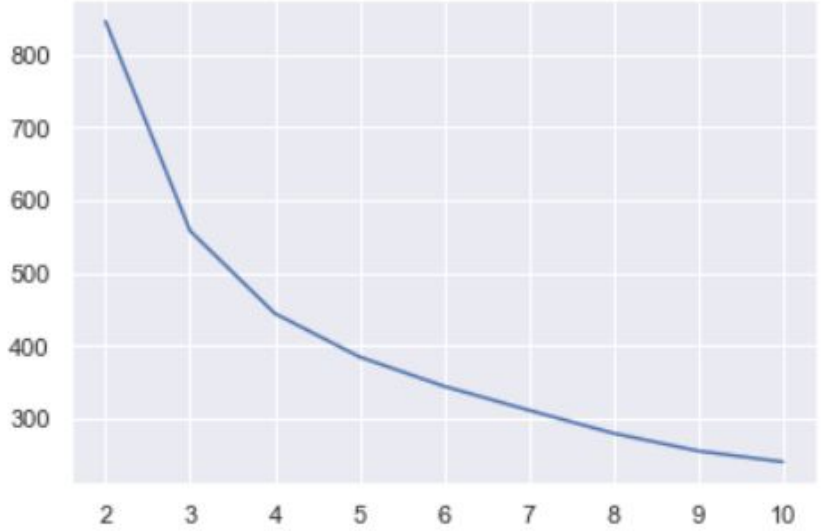


- **gdpp** : Here we see that the graph is on its peak on values between range **0-5000** and then its getting decreased very significantly. Hence, we can understand that the maximum of the countries in this dataset has **low gdpp**.
- **Life_expec** and **total_fer** is having so much ups and downs and not normally distributed.

K-Means Clustering



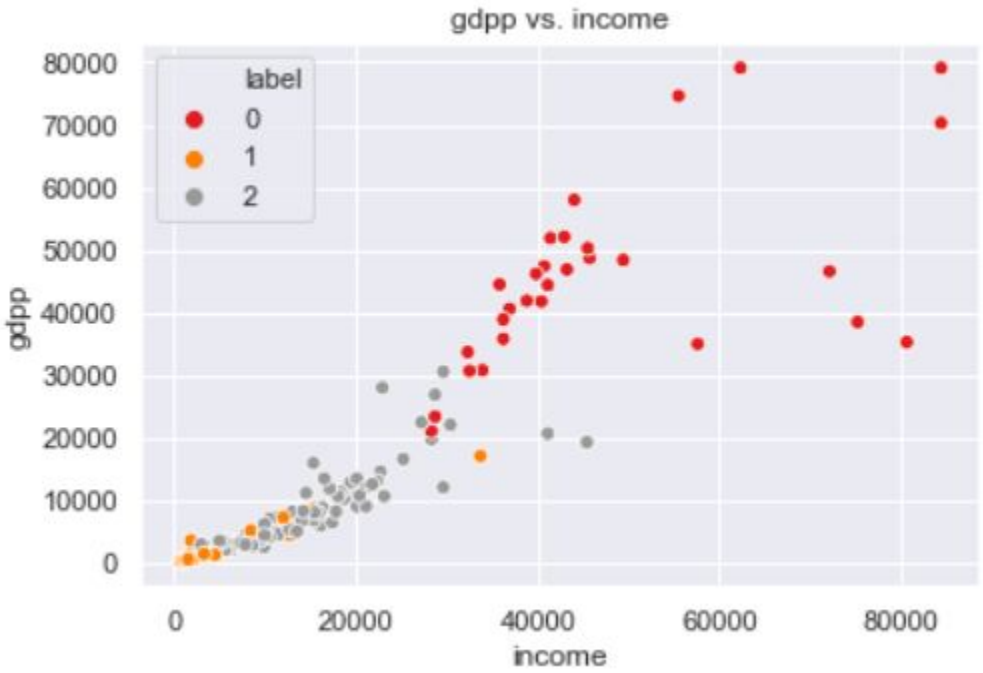
Silhouette Analysis



Elbow - Curve Analysis

- From above observations and noticing the behavior of **silhoutte_score** we can choose cluster values as either **k=3 or k=6**, But ***choosing so many clusters is never a good idea***, So we will go with **k=3** clusters.
- From **elbo-curve** we can see that 3 clusters are getting form. so we will go with **k=3.**

K-Means Clustering contt...

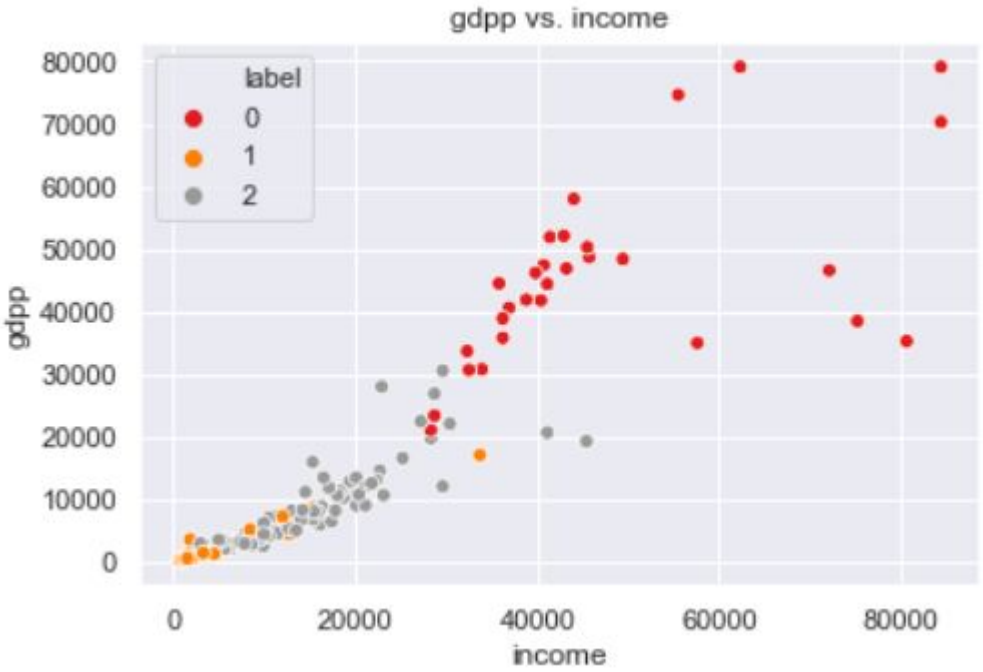


From this plot we can see that we got the 3 clusters with

- **Cluster 0** : **high gdpp** and **high income**
- **Cluster 1** : **low gdpp** and **low income**
- **Cluster 2** : **moderate gdpp** and **moderate income**

This cluster shape is basically in linear form.

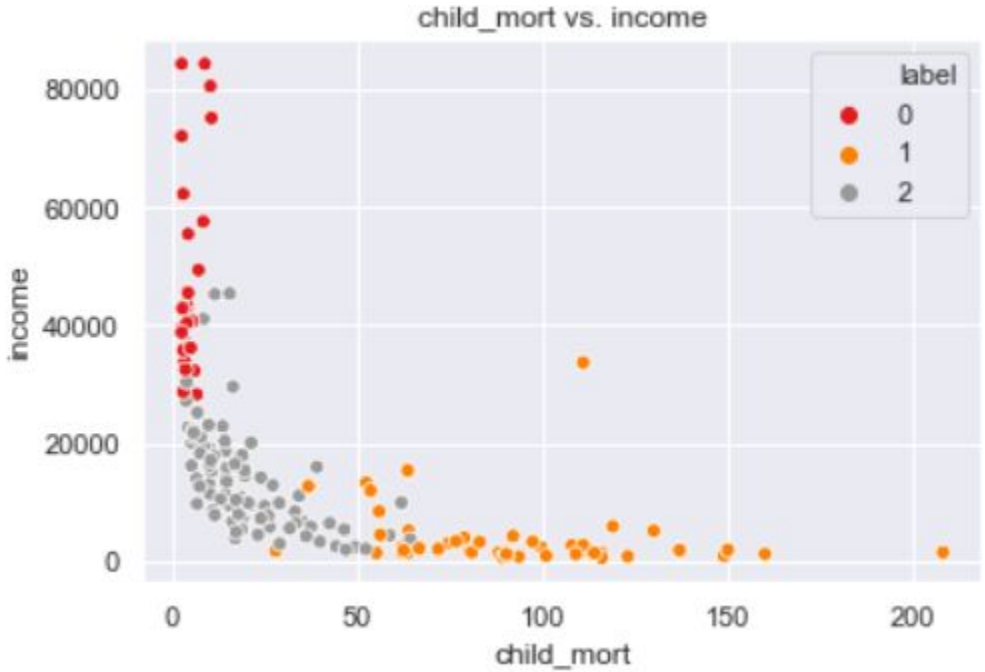
K-Means Clustering contt...



Here we can see we have 3 clusters formed with

- **Cluster 0** : **high gdpp** and **low child_mort**
- **Cluster 1: low gdpp** and **high child_mort**
- **Cluster 2 : low gdpp** and **low child_mort**

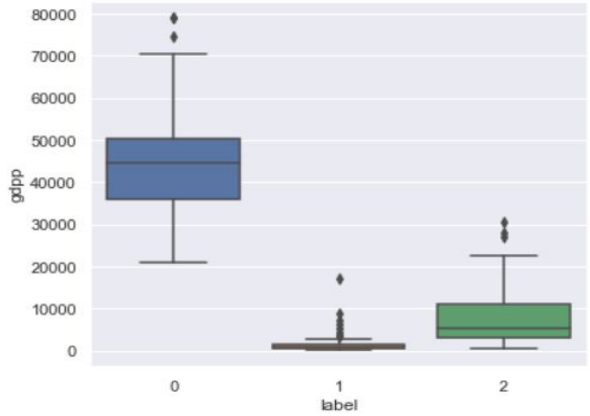
K-Means Clustering contt…

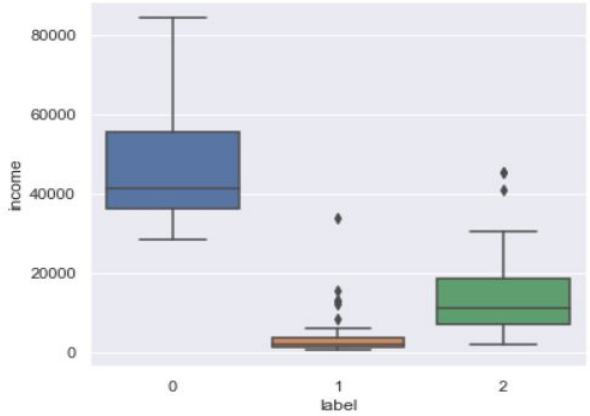


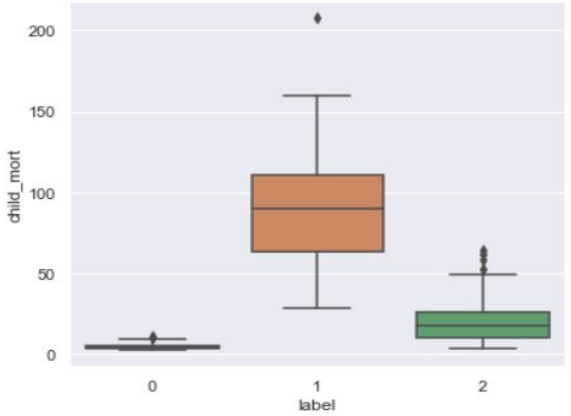
Here we can see we have 3 clusters formed with

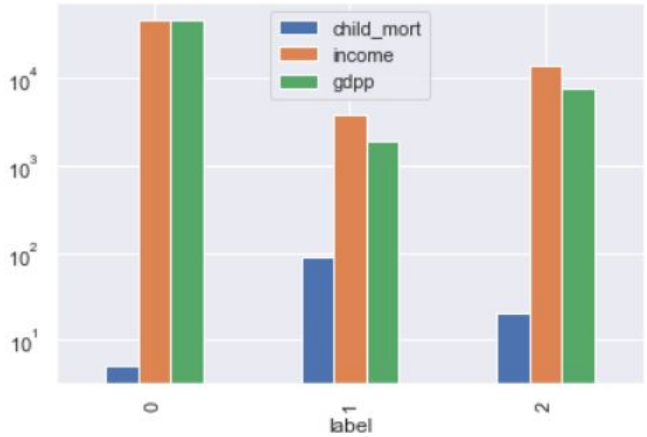
- **Cluster 0: high income** and **low child_mort**
- **Cluster 1: low income** and **high child_mort**
- **Cluster 2: low income** and **low child_mort**

K-Means Clustering contt...









- From graph we can see that..
  - ➔ **Cluster 0:** Developed countries.
  - ➔ **Cluster 1:** Under Develop countries.
  - ➔ **Cluster 2:** Developing countries.
- Here we can see that our main concern is **cluster 1** as this cluster is having **low gdpp, low income** and **high child_mort** values.
- So, we will filter out top 5 countries from **cluster 0** which will be in serious need of aid

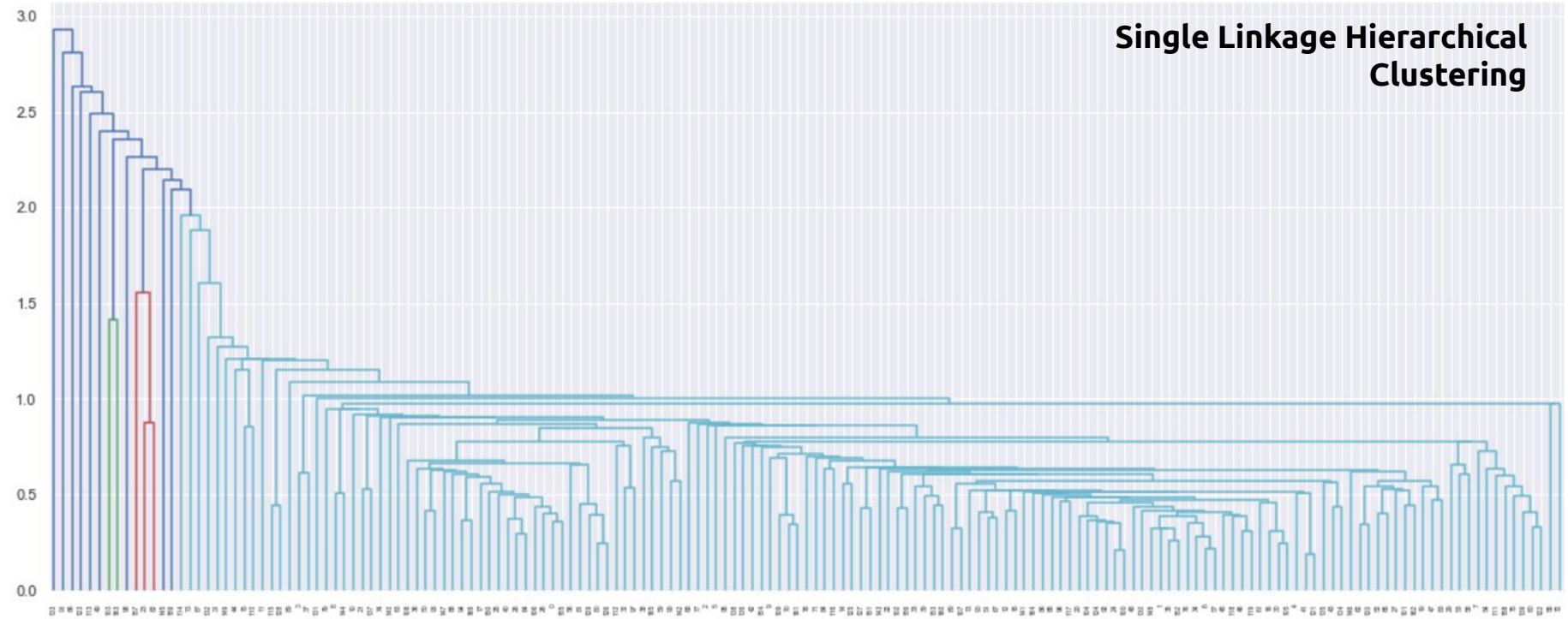
Top 5 Countries from K-Means

These countries are our main point of concern due to - [under-developed-countries]

1. Low gdpp
2. Low income
3. High child_mort

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	label
1	Congo, Dem. Rep.	116.0	137.2740	26.4194	165.664	609.0	20.80	57.5	6.54	334.0	1
2	Liberia	89.3	62.4570	38.5860	302.802	700.0	5.47	60.8	5.02	327.0	1
3	Burundi	93.6	20.6052	26.7960	90.552	764.0	12.30	57.7	6.26	231.0	1
4	Niger	123.0	77.2560	17.9568	170.868	814.0	2.55	58.8	7.49	348.0	1
5	Central African Republic	149.0	52.6280	17.7508	118.190	888.0	2.01	47.5	5.21	446.0	1

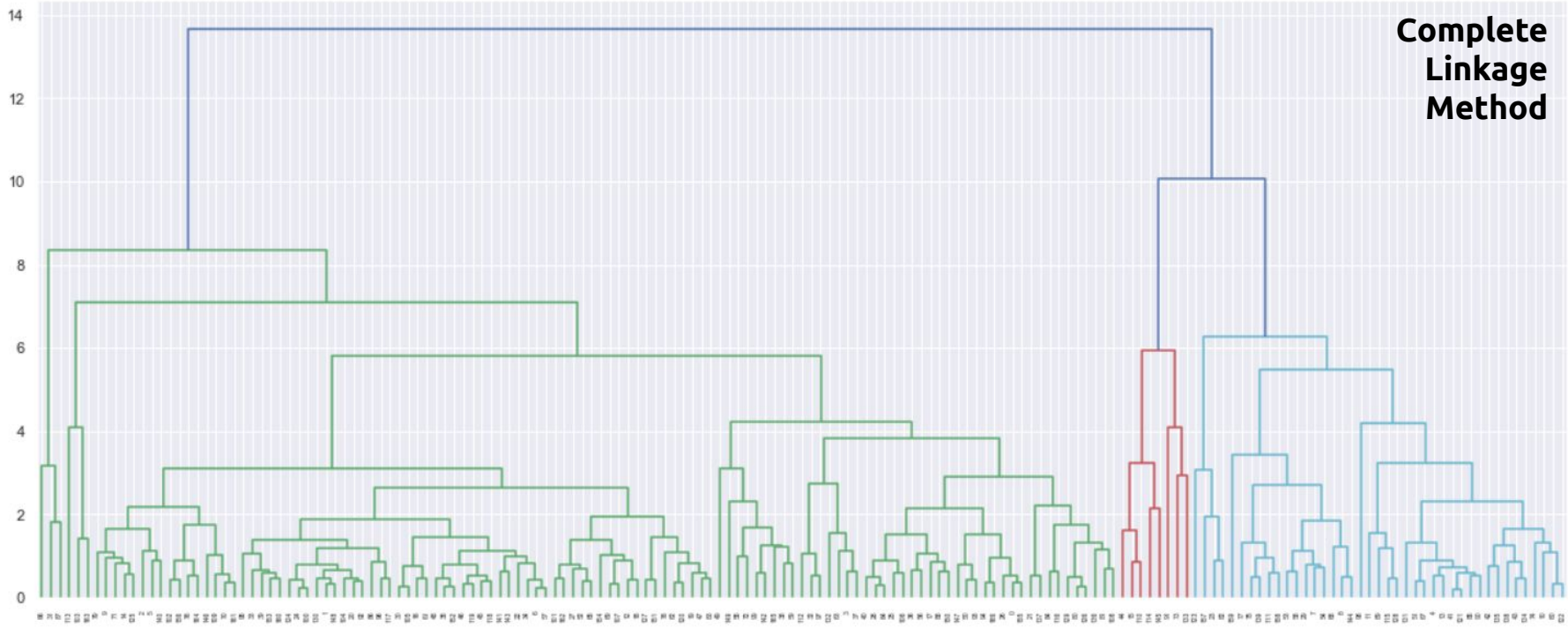
Hierarchical Clustering



From above **single linkage** graph we can see that the clusters are formed are not very good. Its bit confusing to choose the cluster cut-off point. But we can clearly see that the **light blue cluster** are having high majority of countries and then **green and red clusters** are having low majority of countries

So, from above graph it is clear that we are not able to assign the proper cutoff value and choose cluster properly.*

Hierarchical Clustering

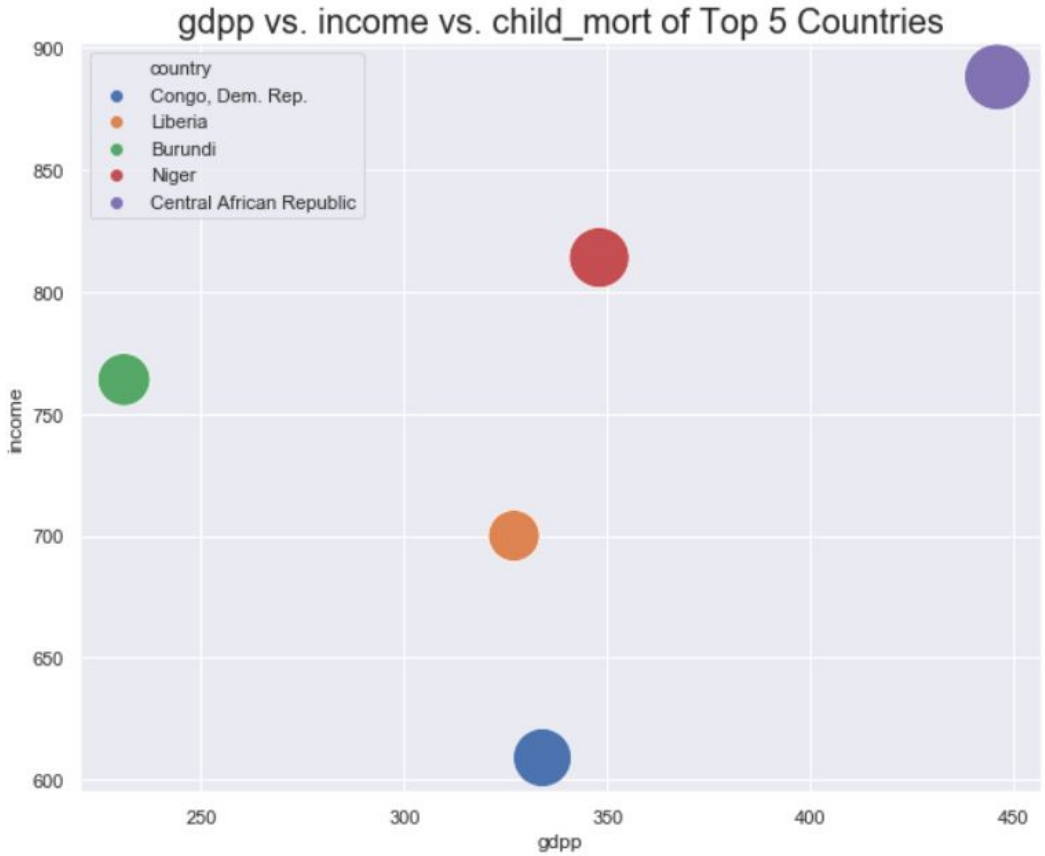


From above **Complete Linkage** graph we can see that at **range 10** the clusters are getting divided into **3 clusters** - **green**, **red** and **sea blue**
green cluster has highest number of countries where as **red cluster** has lowest count of countries.*

Top 5 Countries from Hierarchical Clustering

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	label
37	Congo, Dem. Rep.	116.0	137.2740	26.4194	165.664	609.0	20.80	57.5	6.54	334.0	1
88	Liberia	89.3	62.4570	38.5860	302.802	700.0	5.47	60.8	5.02	327.0	1
26	Burundi	93.6	20.6052	26.7960	90.552	764.0	12.30	57.7	6.26	231.0	1
112	Niger	123.0	77.2560	17.9568	170.868	814.0	2.55	58.8	7.49	348.0	1
31	Central African Republic	149.0	52.6280	17.7508	118.190	888.0	2.01	47.5	5.21	446.0	1

Top 5 Countries story in one graph



In above graph you can see the overall final status of the top 5 countries who are in need of AID.

- **`x-axis` shows `gdpp`**
- **`y-axis` shows `income`**
- **`circles` shows `countries`**
- **`circle-size` shows `child_mort`.** i.e the bigger the circle size, the high is child_mort

Thank You