Help International NGO
Clustering & PCA
Case Study

By

Pravin Pawar

Business Objective

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

- How to use newly received $10 million funding strategically and effectively
- Categorize the countries using socio-economic and health factors
- Choose appropriate countries that are in the direst need of aid

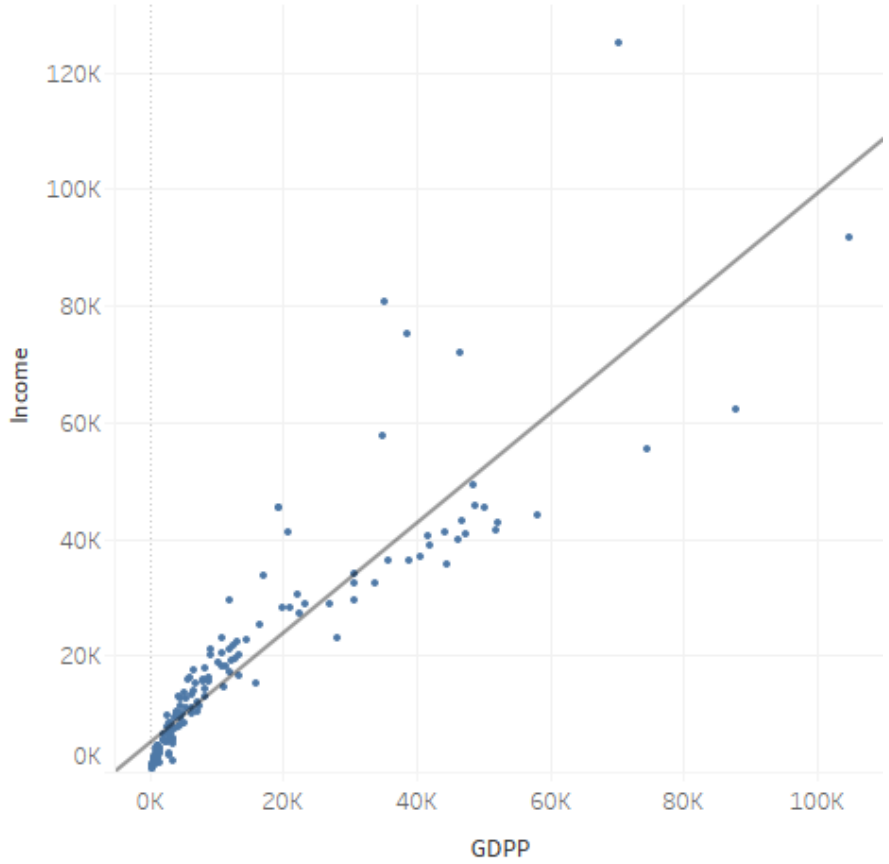
Data Observation & Preparation

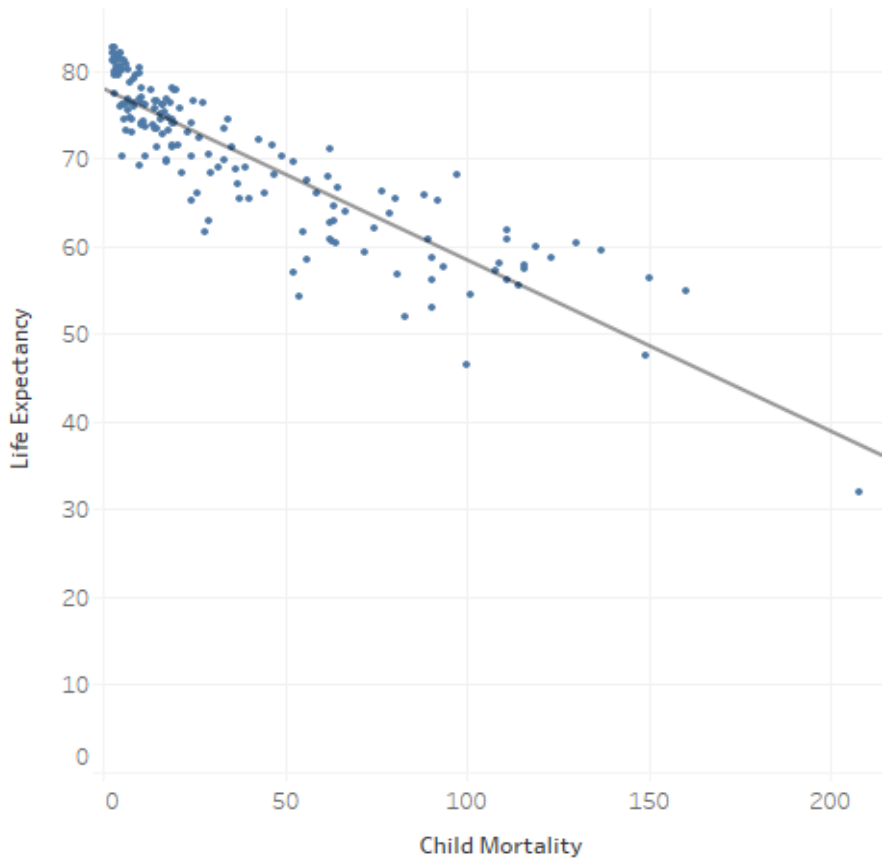
- Total 167 observation found.
- There are no duplicate rows in data set
- Below mentioned features have outlier
  - Income & GDPP have large amount of outliers with very high values from Developed Countries (Can Remove)
  - Apart from that few outlier observation in high value of Child Mortality as well as high inflation rate from under Developed countries (Can't remove as this is our analysis area)
- As mentioned above we have few outliers, but available observation count for analysis are very low (167). So will not remove any outlier.
- Few features are highly Corelated with each other which might cause Multicollinearity issue.
- Will not remove any feature as we are planning to use PCA (Principal Component Analysis), which will take care of multicollinearity issue.

Analysis

Observation

- ❖ Income & GDPP (GDP Per Capita) having linear trend so will consider GDPP for analysis.
- ❖ Similarly Child Mortality & Life Expectancy have liner trend so will consider Child Mortality for analysis.



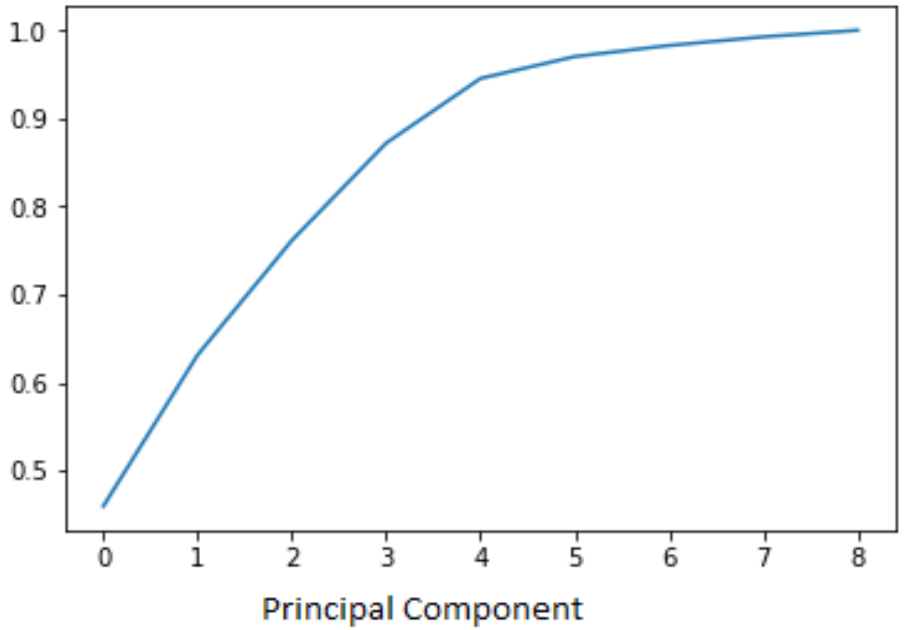


Analysis

Principal Component analysis (PCA)

PCA is used for dimensionality reduction, in our example will reduce our **9** feature variable to **5** principal component by checking cumulative variance explained by component.

Principal Comp	Cum Variance
1	0.46
2	0.63
3	0.76
4	0.87
5	0.94
6	0.97
7	0.98
8	0.99
9	1

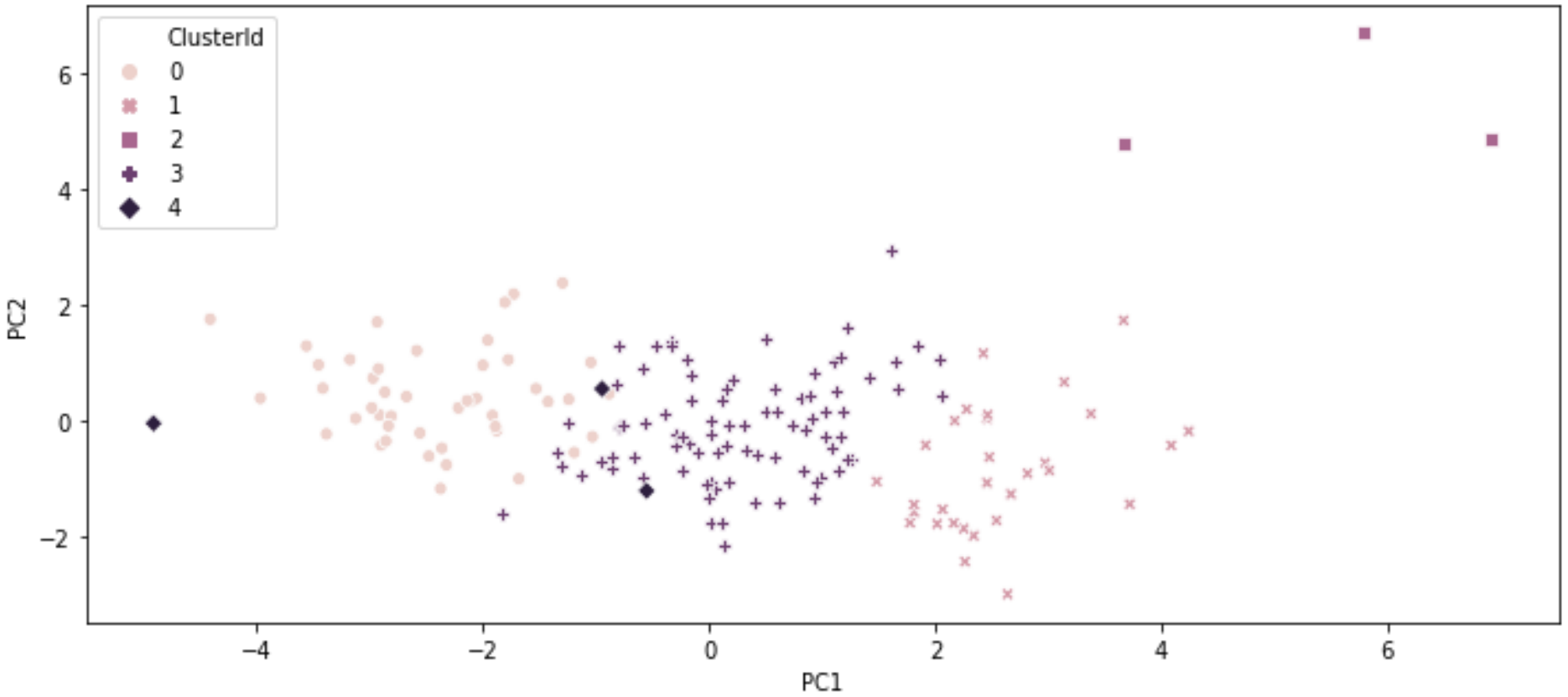


- We can choose 4 or 5 Principal Component based on there Cumulative sum of variance explained.
- In our case we are using **5 principal component** to have maximum variance **(0.94)** while training our algorithm.

Analysis

K-Mean Clustering

- Built machine learning model using K-Mean clustering algorithm
- Using on **Elbow Method (Silhouette Analysis Score)** we divide our data in **5 clusters.**



Cluster	Countries
0	38
1	94
2	30
3	4
4	1

Analysis

K-Mean Clustering

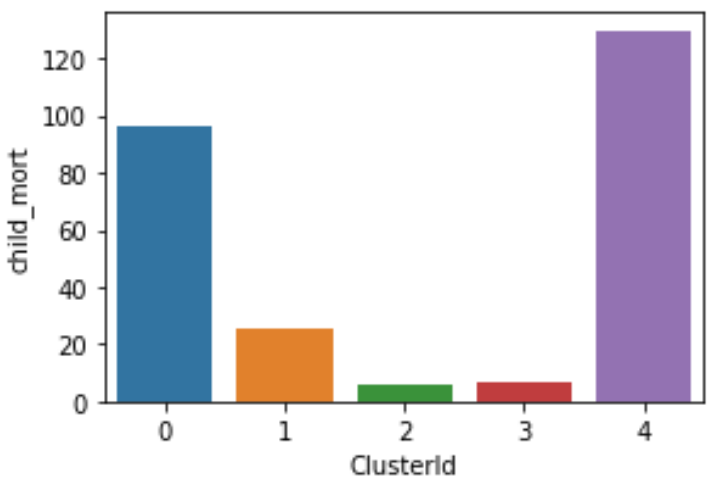
- **Cluster 1 & Cluster 2** are clearly segregated from all other clusters.
- There is some overlap between **Cluster 0 & Cluster 3**
- **Cluster 4** having only 3 data points, but there is some overlap with other clusters.
- Will analyse Clusters against first 2 principal component, which shows good amount of segregation of data.

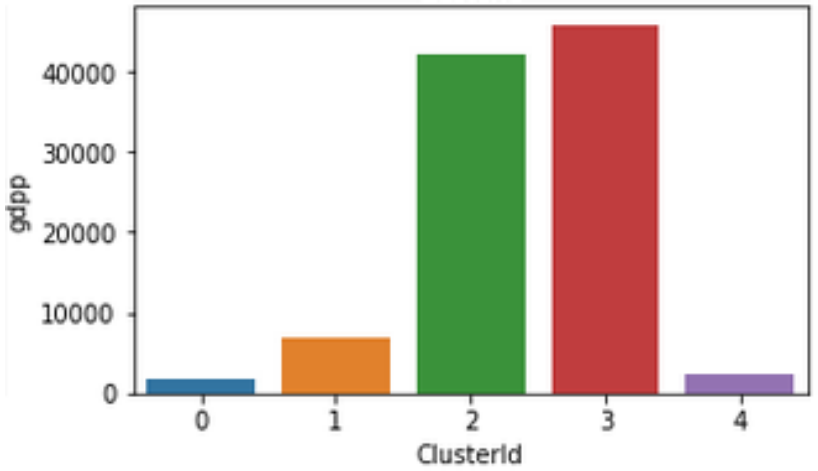
Feature Variable Analysis

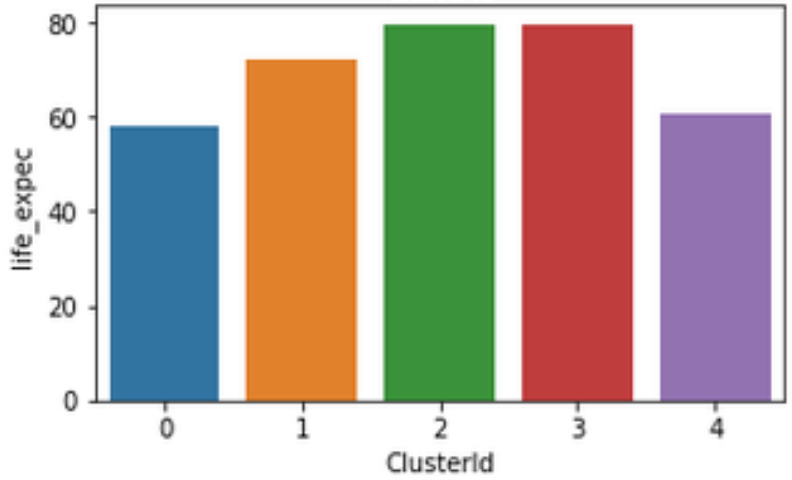
- Features like Child Mortality rate, GDPP, Health Expenditure, Inflation rate, Life expectancy, etc. defines countries growth.
- Country with Low Child Mortality, Low Inflation rate, High GDPP, High Life Expectancy , High Spending on Health related activity are the Developed countries.
- Whereas Under Developed countries normally have High Child Mortality, High Inflation Rate, High Imports, Low GDPP, Low Health Spending, etc.

Analysis

K-Mean Clustering : Feature variable



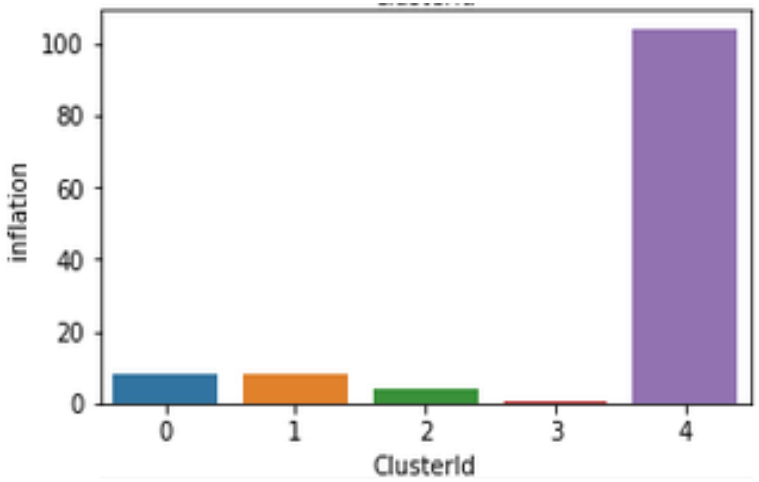


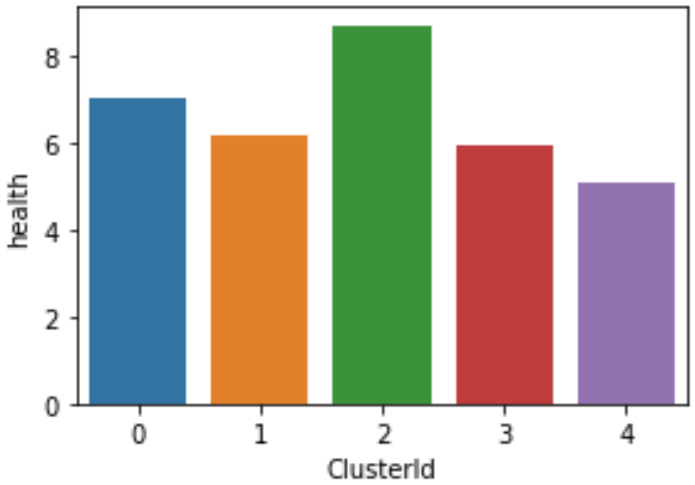


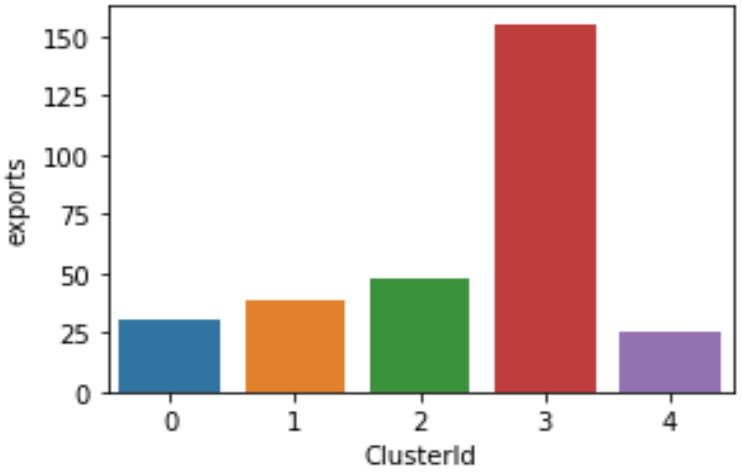
- Cluster 0 & 4 have high mortality rate, very small GDPP, life expectancy is also low as compare to other clusters.
- These two clusters are representing Under Developed Country, whereas Cluster 2 & 3 having very less Child Mortality Rate, High GDPP as well as High Life Expectancy as compare to other clusters these are the Sign of Developed Countries.

Analysis

K-Mean Clustering : Feature variable





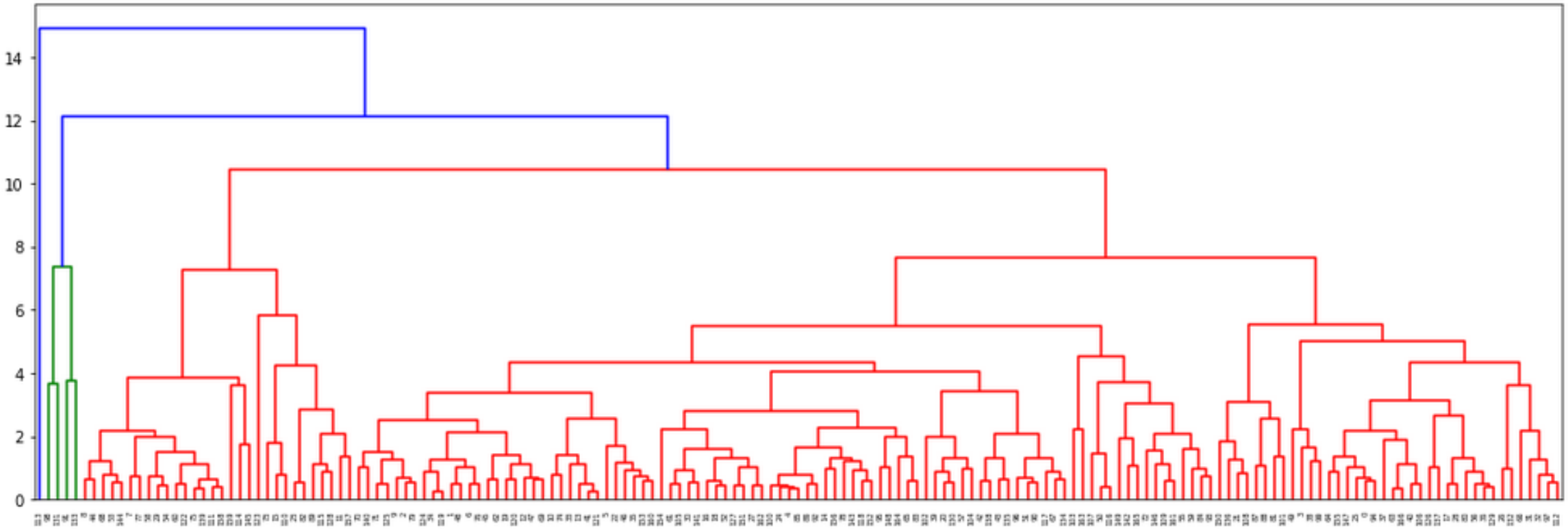


- Cluster 4 is clear outlier in Inflation there is only one Country in this cluster, apart from that Cluster 0 have high inflation rate than Cluster 2 & 3.
- Health Spend is percentage of Total GDP, For Cluster 0 countries GDPP is very low which impacts Health Expenditures.
- Goods Exports ratio is also Low for countries in Clusters 0 & 4, which cause less foreign money inflow.
- **All these features indicates countries from Cluster 0 & 4 are Under Developed countries which are in direst need of aid.**

Analysis

Hierarchical Clustering

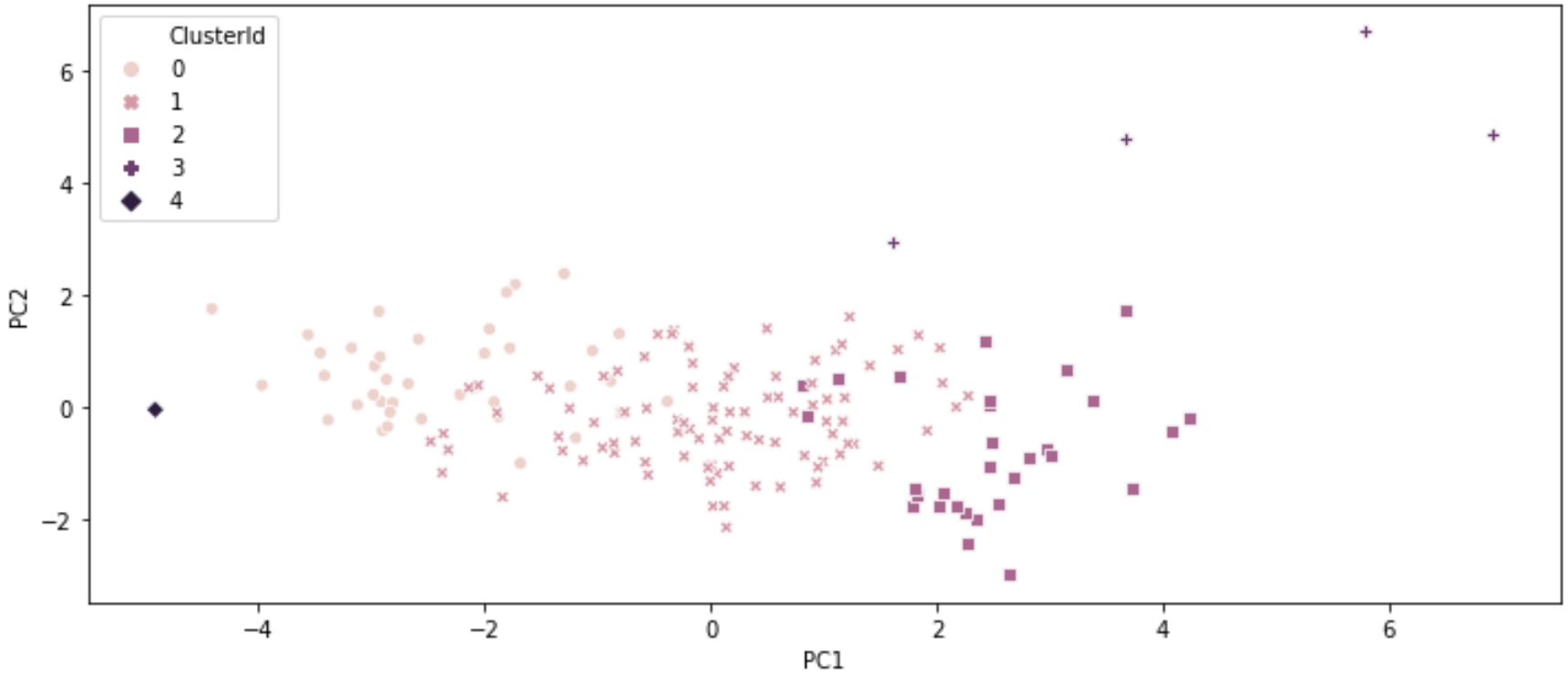
- Built machine learning model using Hierarchical clustering algorithm



- Created **5 clusters** in **Hierarchical Clustering** Model**.**

Analysis

Hierarchical Clustering



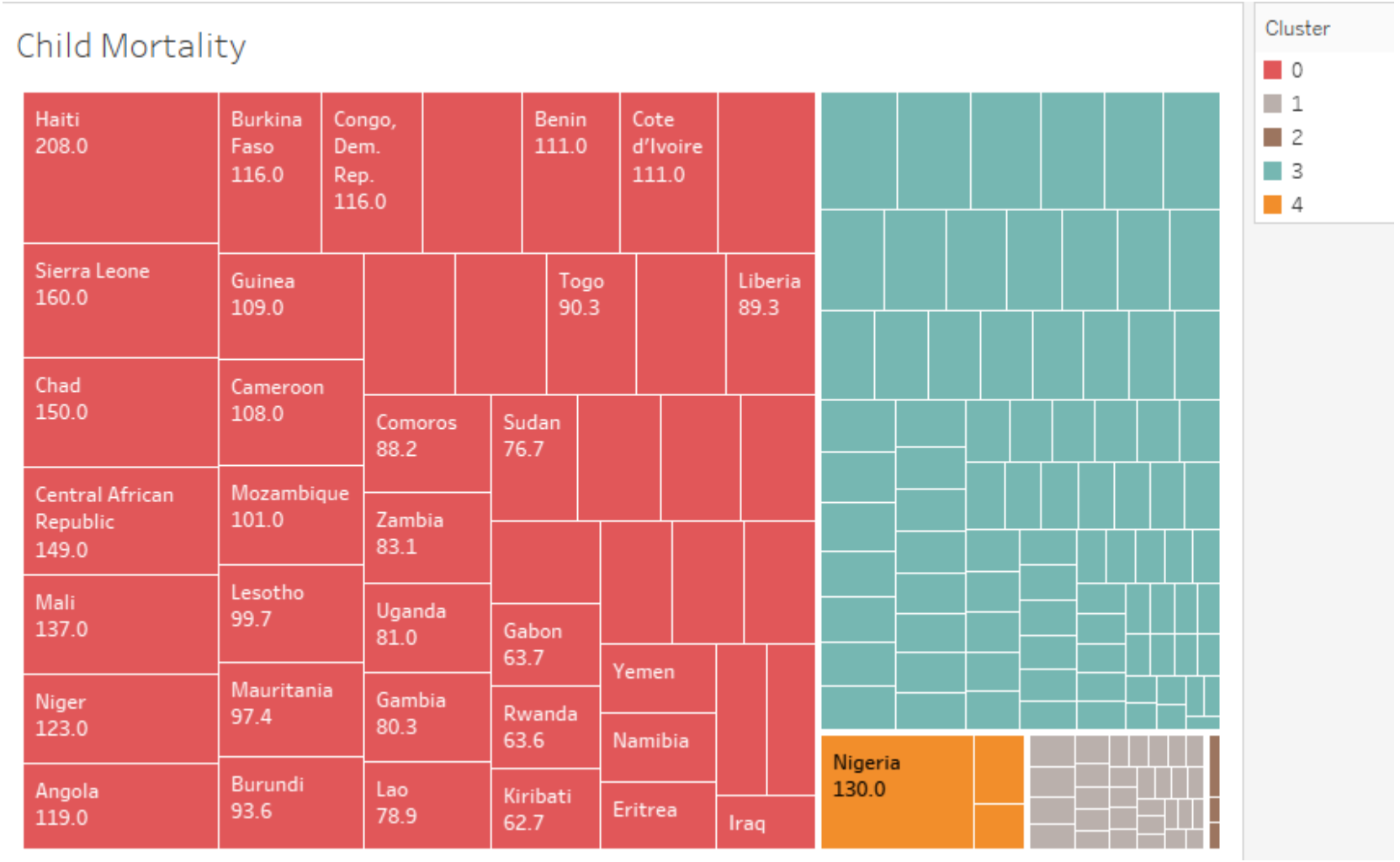
Cluster	Countries
0	38
1	94
2	30
3	4
4	1

- K-Mean & Hierarchical Clustering Model generated similar kind of groups.
- Feature variable analysis also shows same trend like K-Mean, where **Cluster 0 & 4** represents **Under Developed Countries**

Analysis

Cluster 0 & Cluster 4 Analysis : Child Mortality

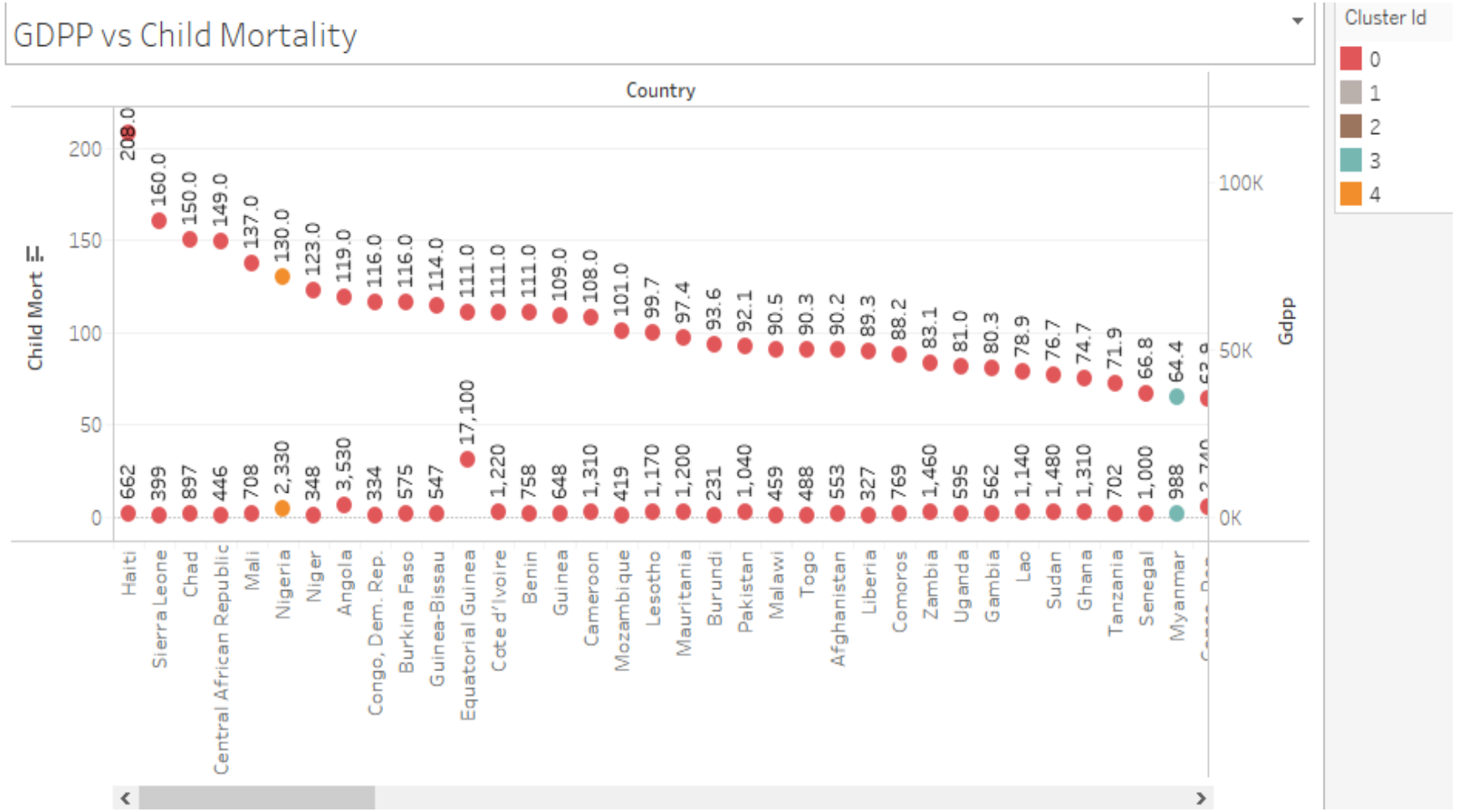
- Cluster 0 have almost all Countries with High Child Mortality problem.
- Cluster 4 also have one country with High Child Mortality Rate.
- **Haiti** from Cluster 0 have highest Child Mortality Rate **208**
- Followed By other countries from **Cluster 0**:
  - **Sierra Leone : 160**
  - **Chad : 150**
  - **Central African Republic : 149**
  - **Mali : 137**
  - **Niger : 123**
- Country from **Cluster 4:**
  - **Nigeria : 130**



Analysis

Cluster 0 & Cluster 4 Analysis : Child Mortality VS GDPP

- GDPP is nothing but GDP Per Capita. (Total GDP / Total Population)
- If GDP of any country is low as compare to there population than GDPP is obviously low.
- Countries from Cluster 0 & 4 **with high Child Mortality have Very LOW GDPP**
  - **Haiti : 662**
  - **Sierra Leone : 399**
  - **Chad : 897**
  - **Central African Rep : 446**
  - **Mali : 708**
  - **Niger : 348**



Analysis

Cluster 0 & Cluster 4 Analysis : Total Fertility VS Child Mortality

- We have almost similar list of countries with maximum Fertility ratio.
- Higher the fertility ratio may cause High Child Mortality in a condition of low GDPP and Low Health expenditure.
- **Niger** from Cluster 0 have highest Total Fertility **7**
- Followed By other countries from **Cluster 0**:

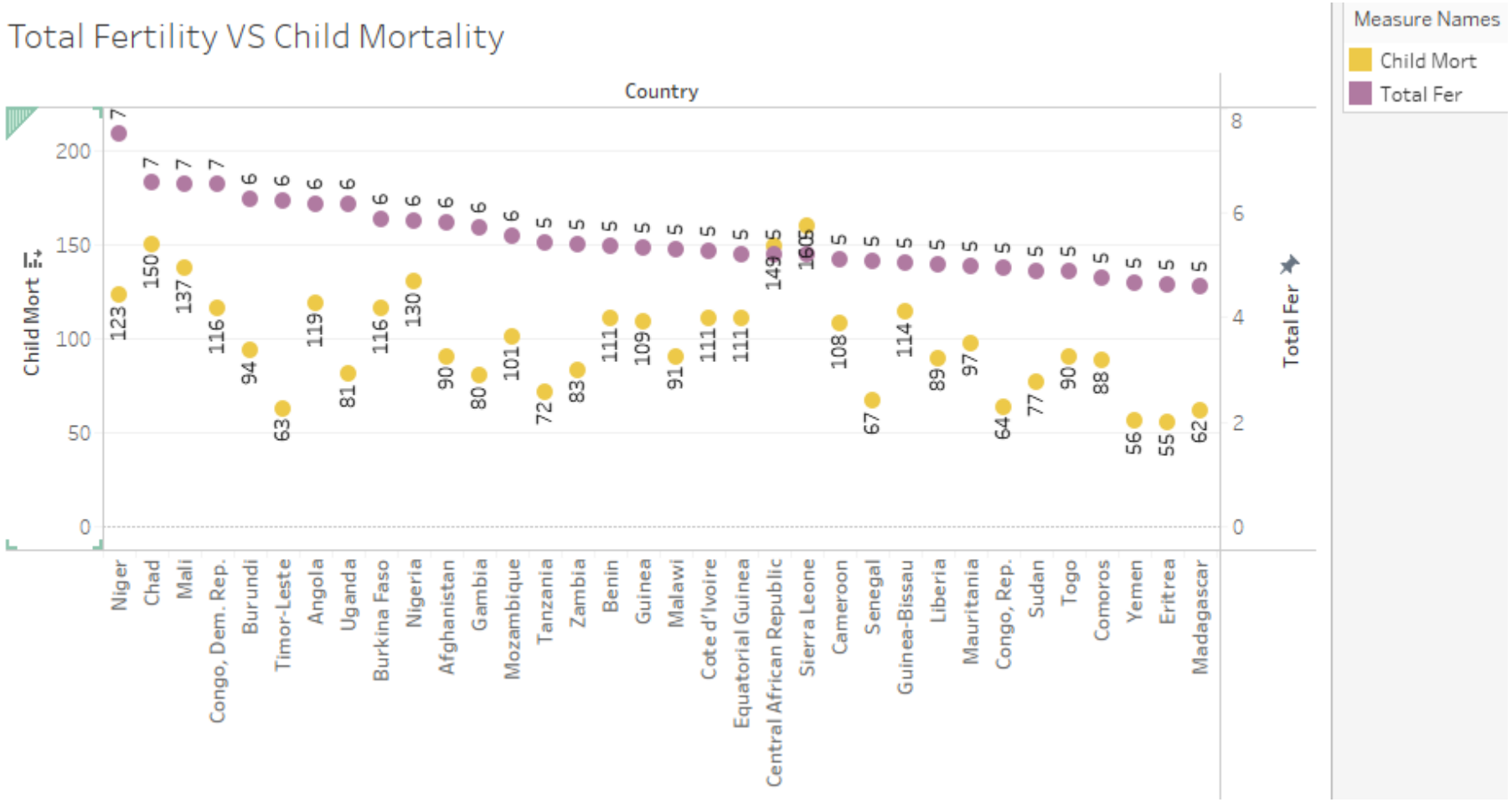
  Chad : 7

  Mali : 7

  Congo, Dem. Rep. : 7

  Burundi : 6
- Country from **Cluster 4:**

  Nigeria : 6



Conclusion

- Key Features which signify countries growth are: **GDPP/Income, Health Expenditure, Child Mortality, Life Expectancy.**
- **HELP International** should target such countries having High Child Mortality, Low Life Expectancy, Low GDPP, Low Health Expenditure, High Inflation, etc.
- Such Countries are listed under **Cluster 0 & Cluster 4.**
- Can choose worst performing countries from these clusters based on above features.
- Few of the recommended countries which are in direst need of aid are:

Sr No.	Country	Child Mortality	GDPP	Health Exp	Inflation	Life Expectancy	Total Fertility
1	Haiti	208	662	7%	5%	32	3
2	Sierra Leone	160	399	13%	17%	55	5
3	Chad	150	897	5%	6%	57	7
4	Central African Republic	149	446	4%	2%	48	5
5	Nigeria	130	2330	5%	104%	61	6
6	Mali	137	708	5%	4%	60	7
7	Niger	123	348	5%	3%	59	7