Clustering and PCA Assignment

SUBMISSION

by

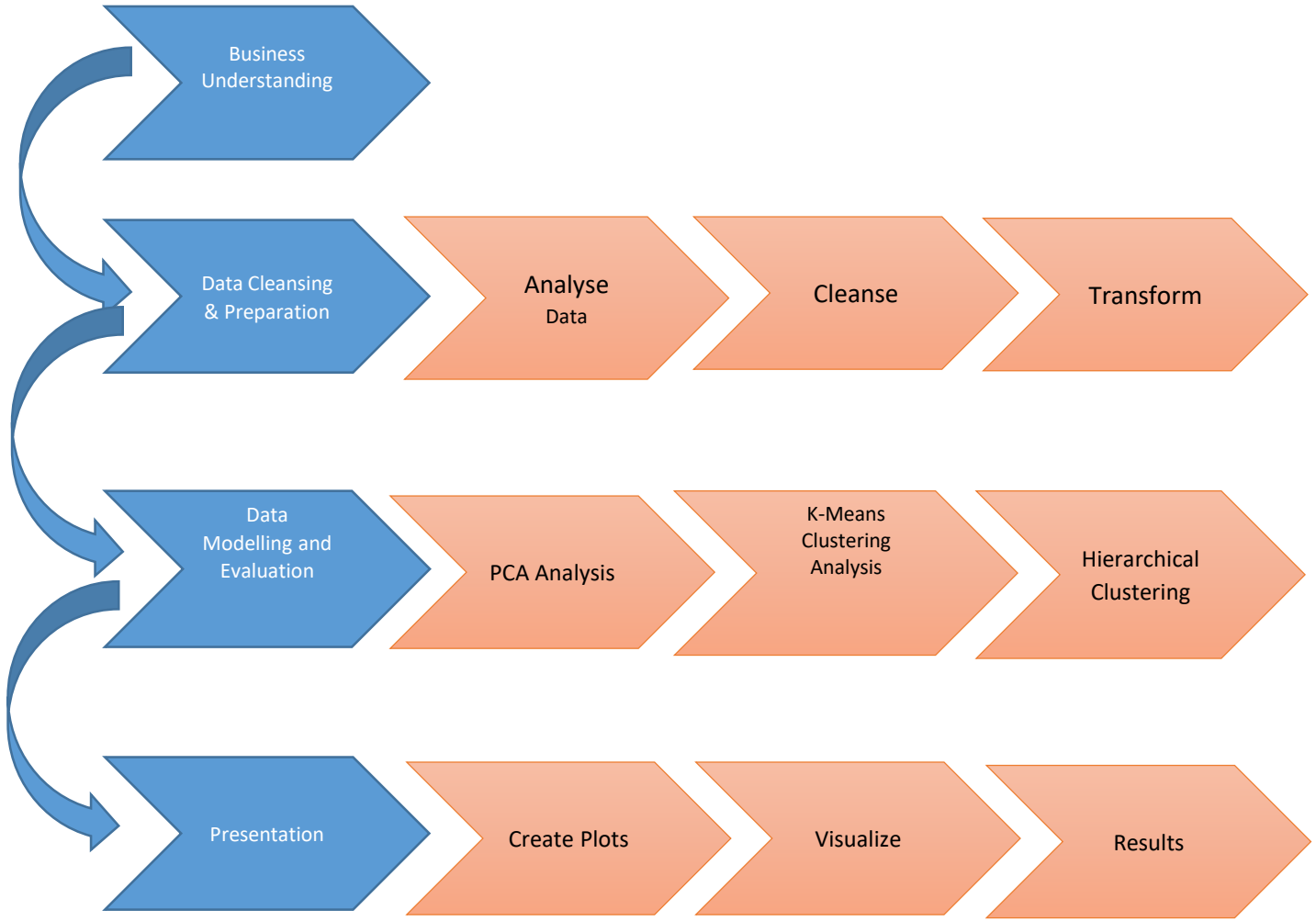
Srinivasan Gopalakrishnan

Date of submission: 24-Feb-2019

Problem Statement

- An NGO organisation;
  - To fund 10 million USD to the countries that are direst need of aid.
  - To Strategically and effectively invest in projects for the backward countries
  - To invest in countries which are poor and/or need relief during the time of disaster and natural calamities.
  - To categorize countries based on Socio-economic factors that determine overall development of the country such as;
    - Income
    - Life expectancy
    - Child mortality
    - GDP
    - Exports & Imports etc.,
  - Identify countries who are in real need of aid so that NGO can fund

Problem Solving Methodology



1] Understand the business context of the problem on the factors driving identifying countries for investment.
2] Assimilating the variables in the data file and understanding through data dictionary. Cleanse the data and transform data for modelling.
3] Develop PCA (Feature Reduction) Analysis and Conduct K-Means Analysis on the Data to identify K clusters.
4] Develop Hierarchical Model and identify number of clusters.
5] Generate Insight on group of countries for investing and present details visually.

Assumptions

- Data provided includes outliers
  - The data contains outliers in almost all columns
  - Since the number of data provided is very small, the outliers data is imputed with values using Capping and Flooring methodology.
- The Data sample provided is enough to get some insight.

Data Cleansing

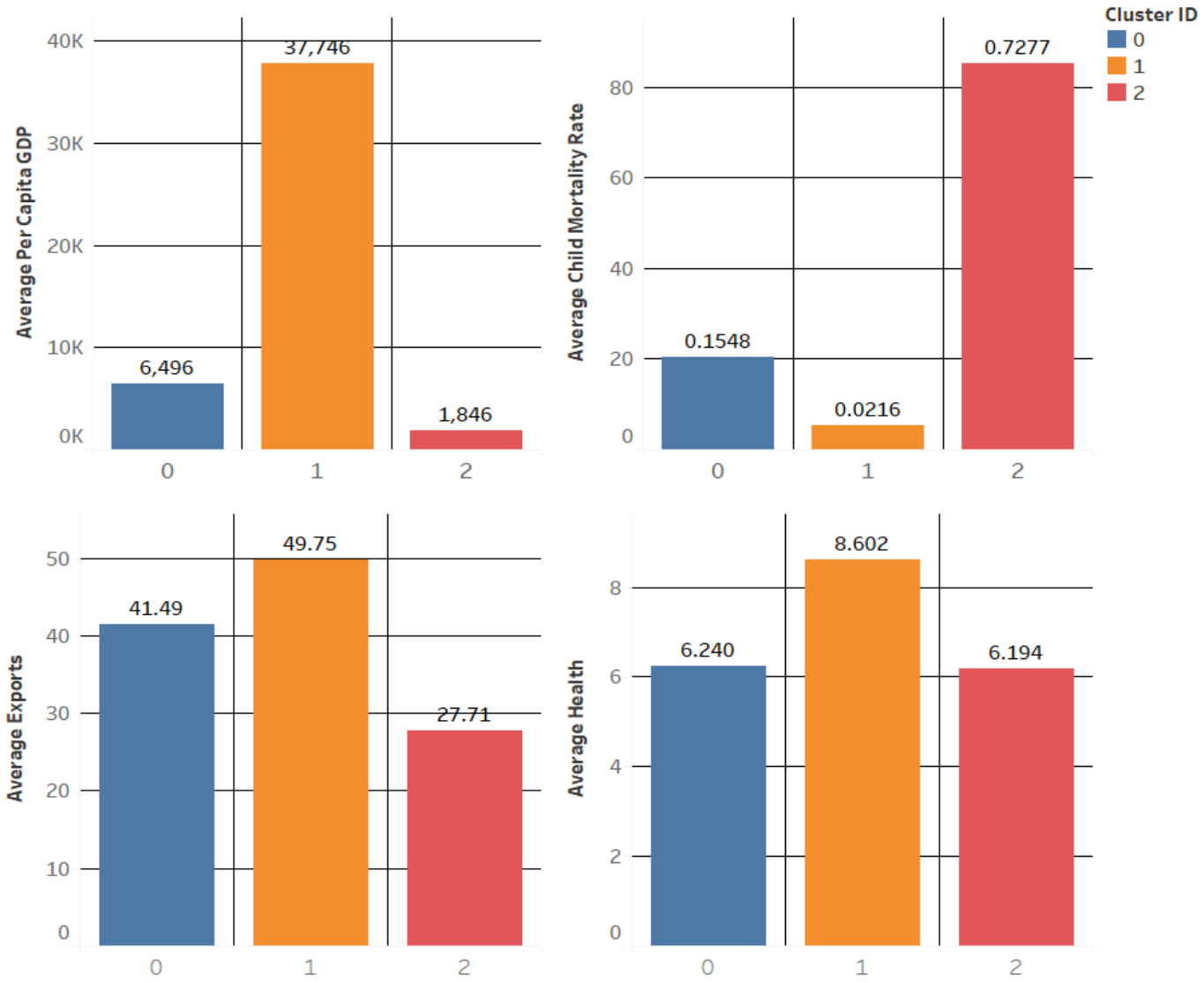
- Data contains no null values or duplicates
- Convert data type of income and GDPP from integer to float
- There are 167 rows 10 columns in the data set
- All the numerical variables have outliers
- Due to small size of data sample, the outliers are imputed with values using capping and floor methodology.(0.01 to 95% scale).
- Data has been standardized using StandardScaler of skelearn library

Data Modelling & Evaluation

- PCA Analysis
  - PCA Analysis carried out on the data set in order to obtain the Principal components
  - The number of features or Dimension reduced to 5, which explains as high as 95% of the variance in the data
  - Model evaluated using Scree plot and analysing Explain variance Ratio and Cumulative Variance Ratio
  - Plot Correlation Matrix verify the collinearity between featured variables
  - Scatter Plot to verify the Principal components projections.
- HOPKINS Statistics
  - Hopkins Statistics of ~.70 indicates very high tendency for clustering
- K-Means Clustering
  - Elbow Analysis
  - K-Means cluster (based on PCA) with K=6 and K=3
- Hierarchical Clustering

Presentation

- Different Plots are created in the below order
  - Bar Plots for distribution of Socio-Economic Factors across clusters (0,1 and 2).
    - Cluster wise Average Per Capita GDP
    - Cluster wise for Average Child Mortality rate
    - Cluster Average Exports
    - Cluster wise Average Health
    - Cluster wise Average Income
    - Cluster wise Average Life Expectancy
  - Few Scatter plots showing few socio economic factors which key are influencers in strategic making of funding across countries. It includes ;
    - Cluster wise impact of Net Income against Child Mortality Rate
    - Cluster wise impact of Total Fertility Rate against Child Mortality Rate
    - Cluster wise impact of Net Income against Life Expectancy
    - Cluster wise impact of Net income against Total Fertility Rate
    - Cluster wise impact of inflation on GDPP



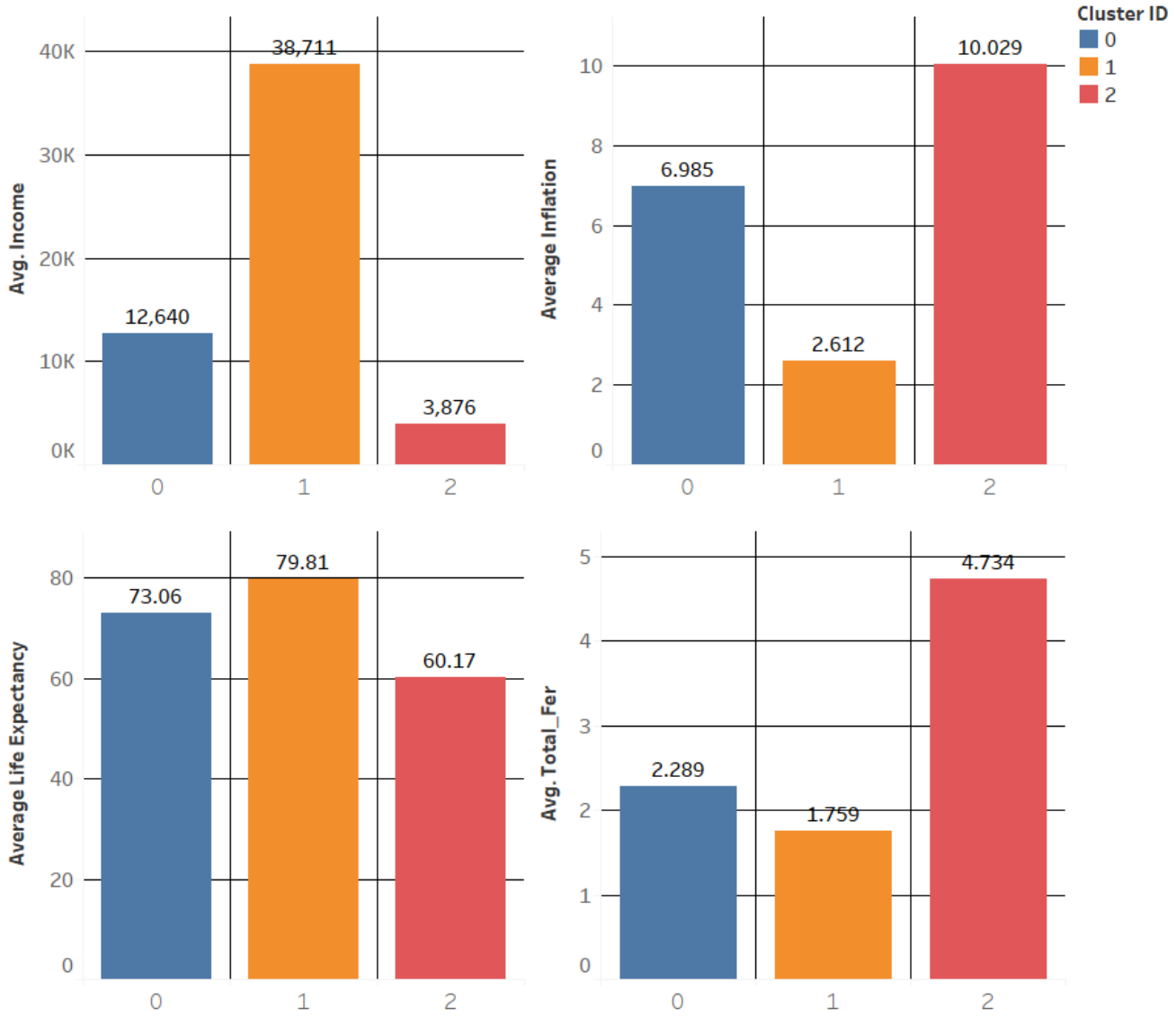
1]

From bar plots, the Cluster 2 has the list of strategic countries having key metrics of socio-economic and health factors which are key criteria and should be looked upon for funding.

2]

The cluster 2 has following;

- Low GDP per-capita of **1846**.
- Highest mortality rates of **~85% or ~0.7277** (normalized)
- Below par exports of total GDP **~27.71%**
- Poor health spending of **6.194%**

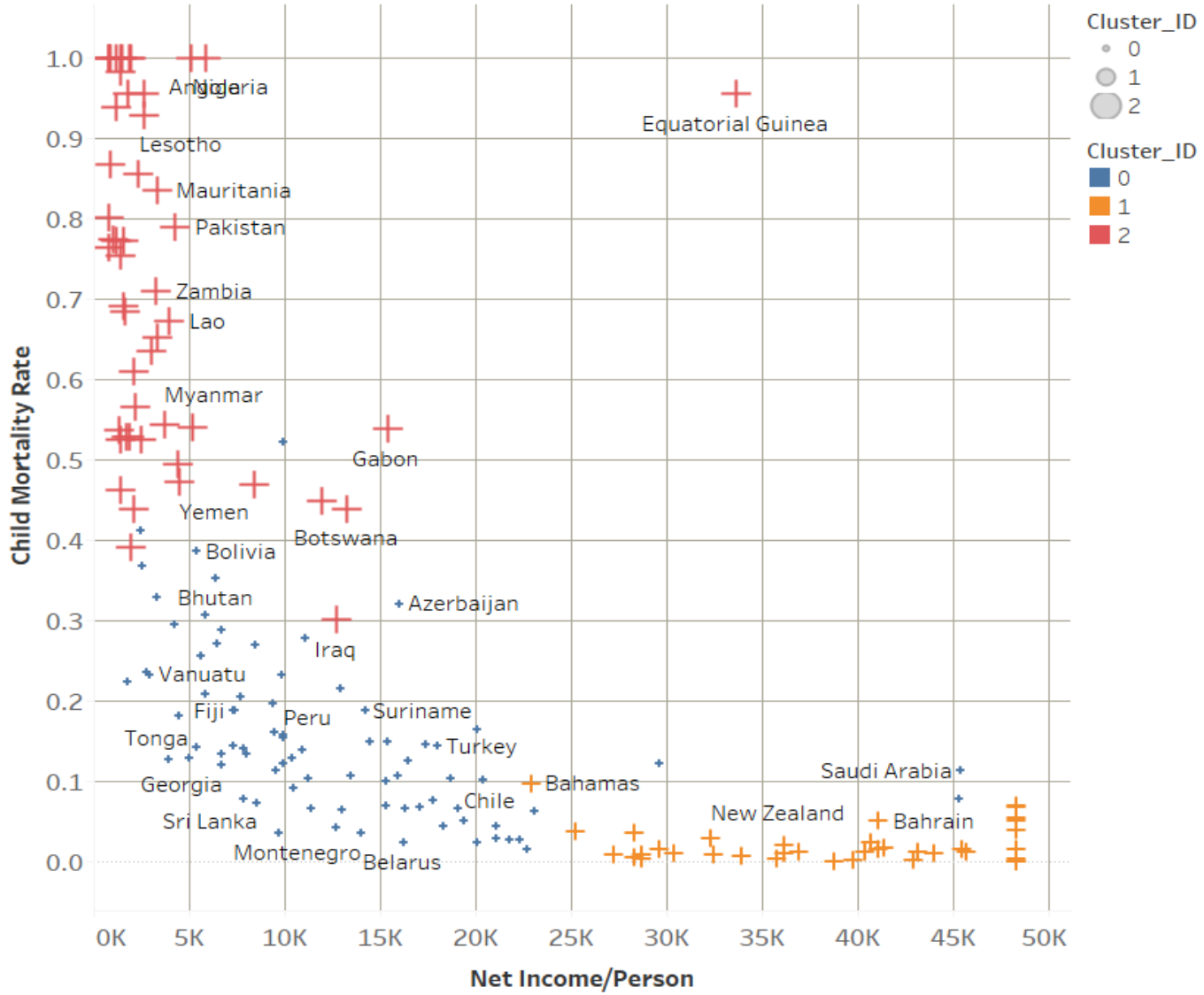


1]

Again, Cluster 2 list of strategic countries which are having following metrics

- Low Average net income per person of **~3876**
- high inflation rate of **~10.03%**
- Low average life expectancy of infants **~60.17%**
- high fertility rate of **~4.73** per woman

Plot-1: Net Income Vs Child Mortality Rate



1]

The scatter plot indicates cluster pattern across countries for Net Income against Child Mortality Rate.

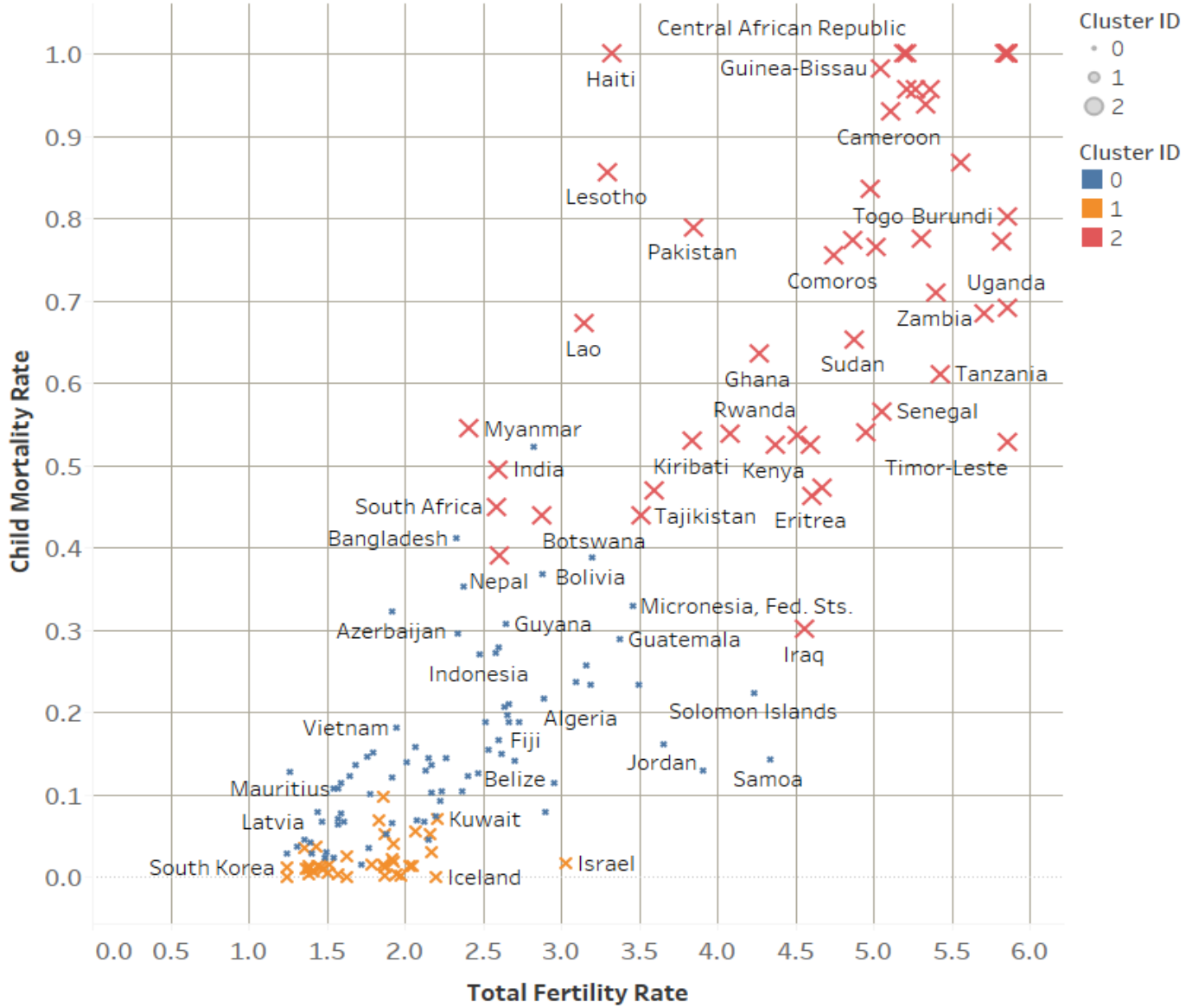
2]

Net Income is inversely proportional to Child Mortality Rate .

3]

It is one of the Key strategic economical factor which influences funding for those poor and backward countries.

Plot-2: Fertility Rate Vs Child Mortality Rate



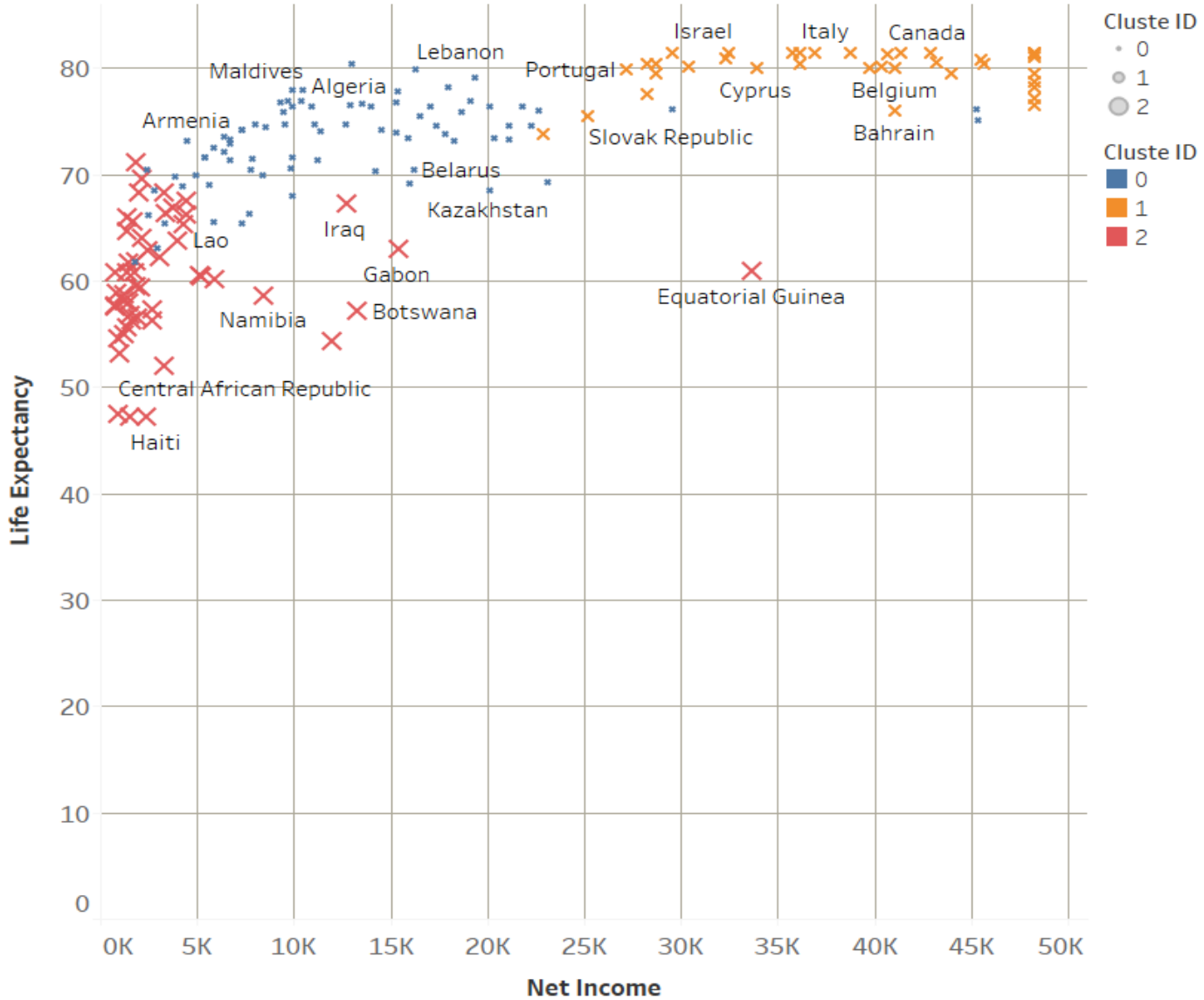
1]

The scatter plot shows the impact of Fertility rate in the third world and African countries are one of the social factor which influences the child mortality rate.

2]

The Child Mortality rate proportion to number of infants in the family

Plot-3: Net Income Vs Life Expectancy



1]

The scatter plot shows the impact Income on Life Expectancy or average life span on each of the countries.

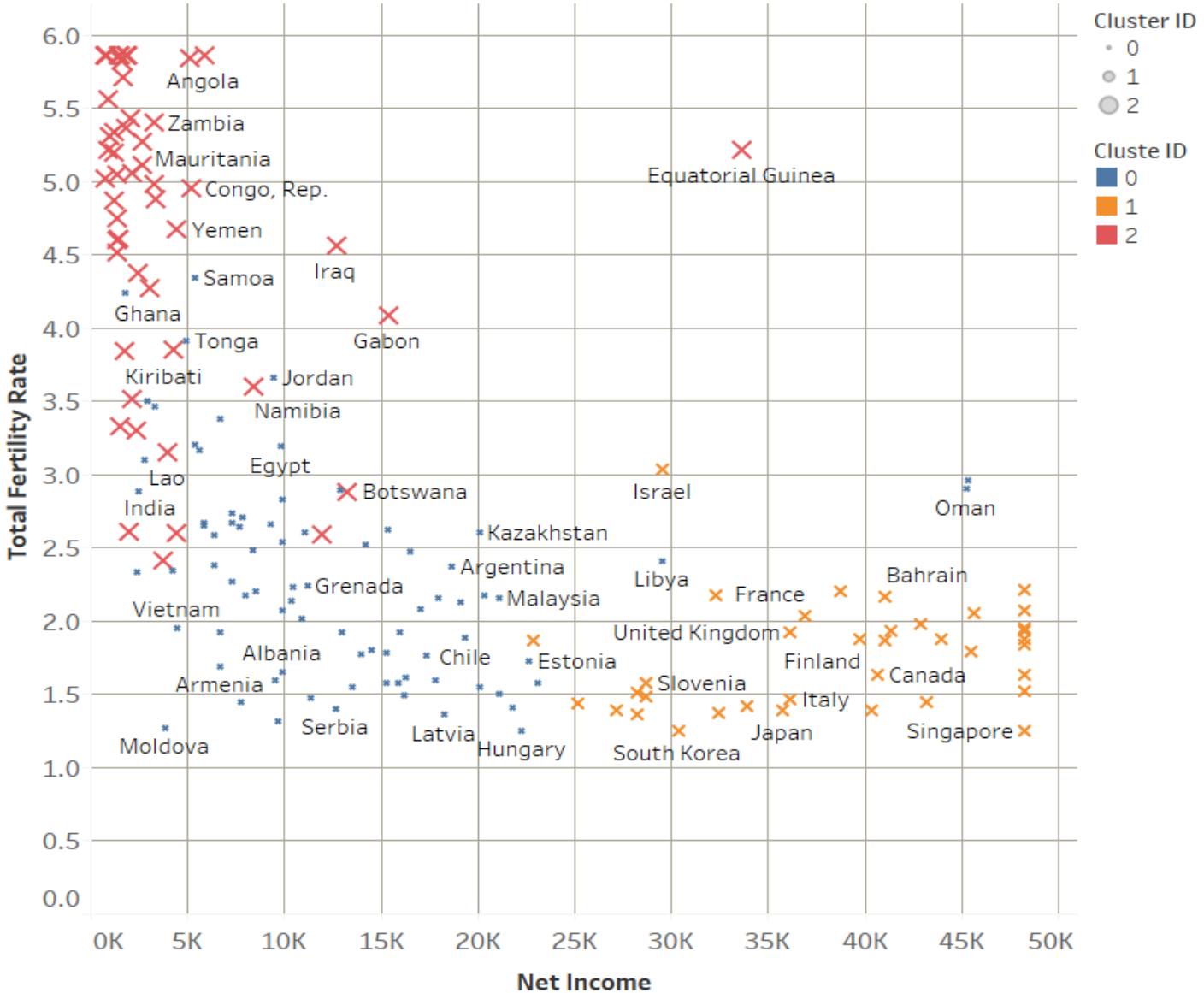
2]

Again, Income is the Key factor influencing the Life Expectancy.

3]

Higher the Income group have higher lifespan having more than 70 years.

Plot-4: Net Income Vs Fertility Rate



1]

Again, the scatter plot shows the impact Income on Fertility Rate.

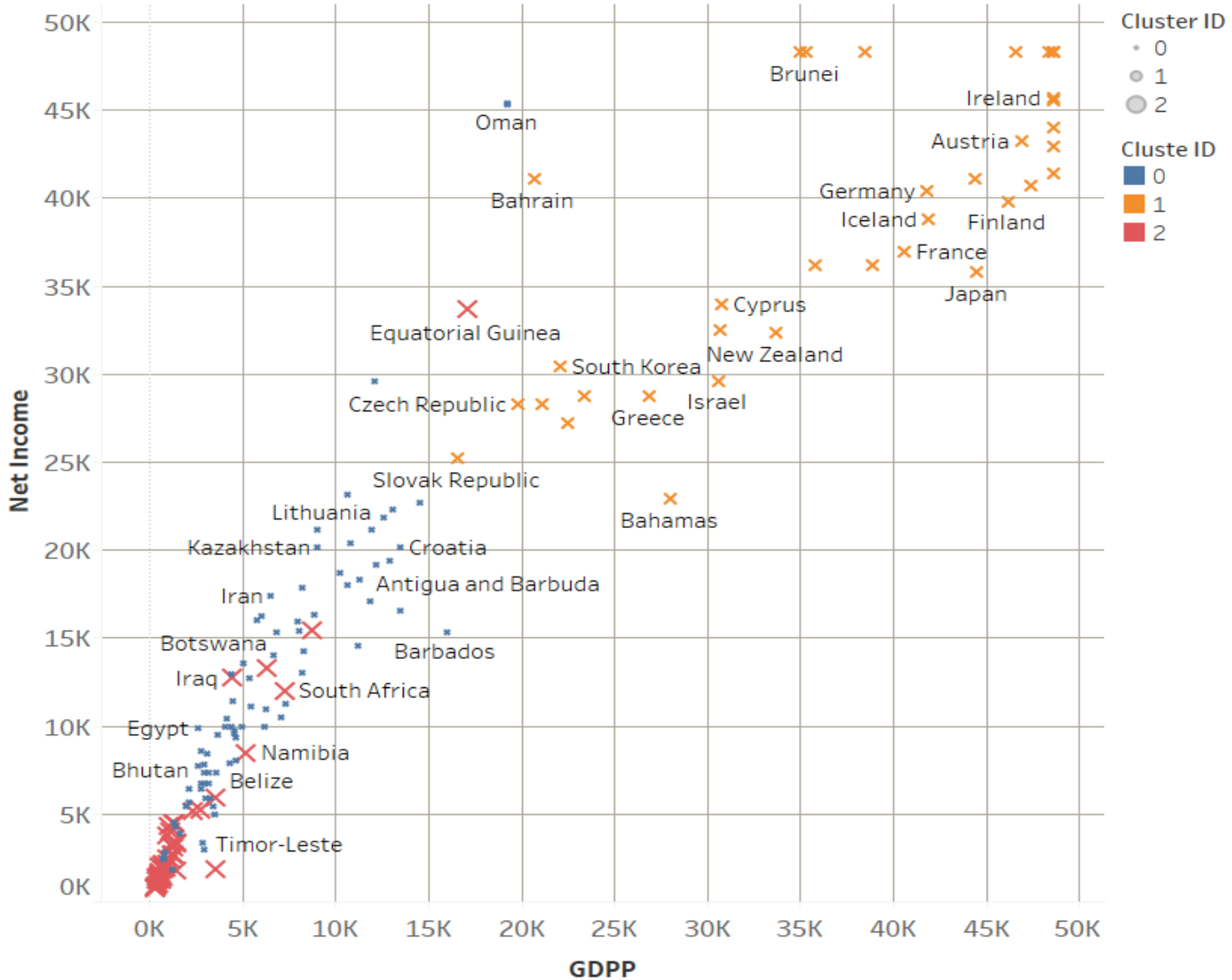
2]

Higher the Income group has lower fertility rate , as having not more than 2 children's.

3]

while the countries which are economically backward and has lower GDPP has more than 5 children.

Plot-5: GDPP Vs Net Income



1]

The scatter shows the how the country GDPP influenced by Net income of individuals and invariably countries growth.

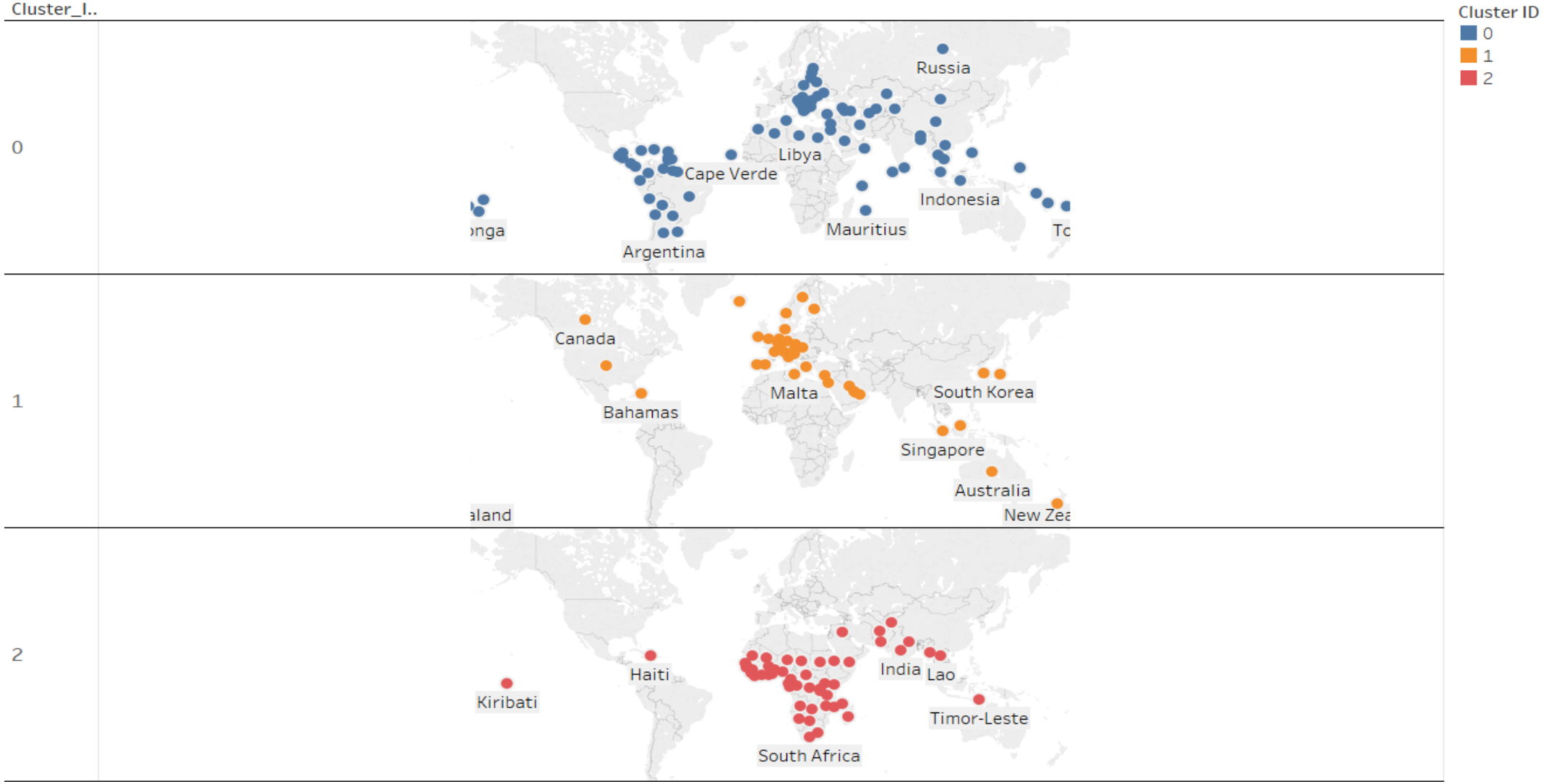
2]

GDPP directly proportional to Net Income which has direct impact on the economic growth of the country.

3]

Another Key factor which needs to be looked in to filtering down the countries where investment should be targeted.

Plot-6: Geo Mapping - Country wise Cluster Classification



Note: *The recommended countries will be the countries for strategic investment for NGO is from Cluster 2 marked by Red circles in Geo Map, where majority of populations either well below poverty line or disaster struck due to political or economical reasons.*

Conclusion -- Recommendation

- Considering HELP NGO Investment strategy of investing 10M USD, is depends on the following key socio economic factors predicted based on the sample data analysis . They are ;
  - Child Mortality Rate
  - Income
  - Total Fertility Rate
  - Health Factors
  - GDPP

- The recommendation of the top 10 destinations would be Angola, Central African Republic, Congo, Chad and some of the poor African countries etc. in the group.