HELP INTERNATIONAL
CLUSTERING AND PCA

SUBMISSION

Student Name:
Divij Jawarani

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

Business Objective: Help International received a funding of $10 million. They would like to use these funds to help countries in dire need of aid. They would like to do this strategically and effectively

Goals of Data Analysis: To HELP identify 5 countries which require aid, and which can he helped most by HELP international

Methodology: The data analyst is to first perform PCA on the dataset to find out principal components in the data. Then use K-means clustering and hierarchal clustering and analyze the socio-economic factors in the data to produce the list of countries to be selected for aid by the company

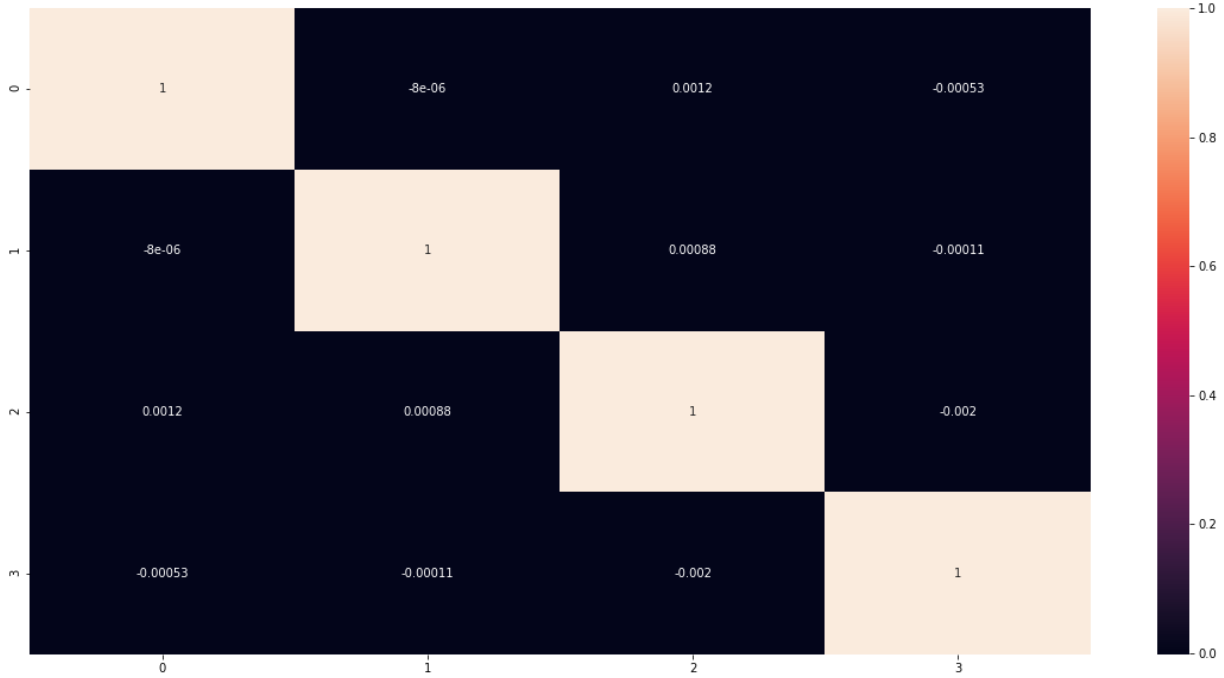
METHODOLOGY

1. Importing the Countries data csv to understand the data at hand
2. Doing outlier analysis
3. Perform PCA on the dataset with selected number of components
4. Create a new dataset with the selected components
5. Perform k-means clustering on the PCA dataset to form clusters
6. Perform hierarchal clustering on the same dataset
7. Visualise the clusters formed by analysing the original variables given
8. Select 5 countries which require the most aid

PRINCIPAL COMPONENT ANALYSIS

Principal component analysis was done on the data for dimension reduction without losing much data.
4 principal components were selected

We can see in the below plot that there is hardly any correlation between the components. Thus clustering can be easily performed



Results of PCA

	PC1	PC2	PC3	PC4	Feature
0	-0.419519	0.192884	-0.029544	0.370653	child_mort
1	0.283897	0.613163	0.144761	0.003091	exports
2	0.150838	-0.243087	-0.596632	0.461897	health
3	0.161482	0.671821	-0.299927	-0.071907	imports
4	0.398441	0.022536	0.301548	0.392159	income
5	-0.193173	-0.008404	0.642520	0.150442	inflation
6	0.425839	-0.222707	0.113919	-0.203797	life_expec
7	-0.403729	0.155233	0.019549	0.378304	total_fer
8	0.392645	-0.046022	0.122977	0.531995	gdpp

CLUSTERING

Two methods of clustering was performed on the PCA dataset

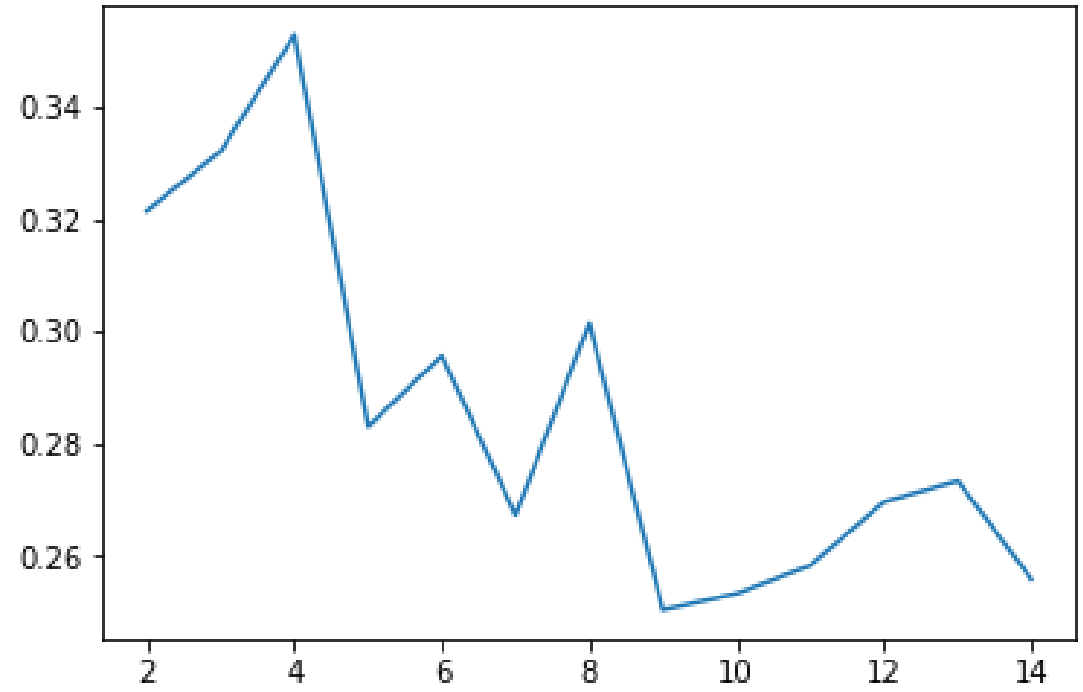
1. K-means clustering
2. Hierarchal clustering

3 clusters were formed by using k-means clustering

2 clusters were formed by using hierarchal clustering

K-MEANS CLUSTERING

Silhouette Analysis results



After conducting silhouette analysis, we selected 3 clusters

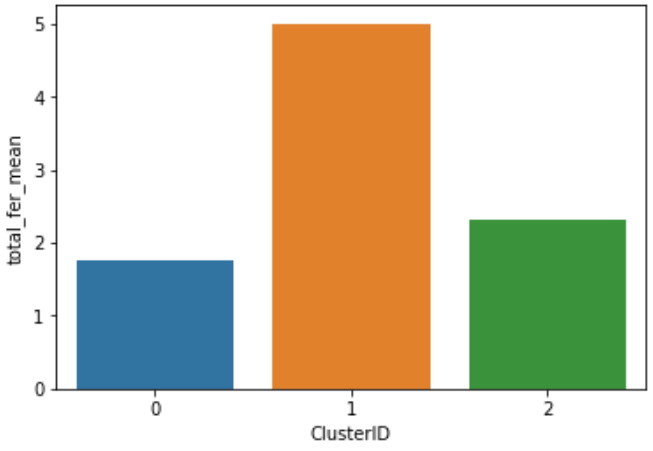
ClusterID	country
0	36
1	47
2	84

- Cluster 0 – Developed countries
- Cluster 1 – Under-developed countries
- Cluster 2 – Developing countries

K-MEANS CLUSTERING - RESULTS

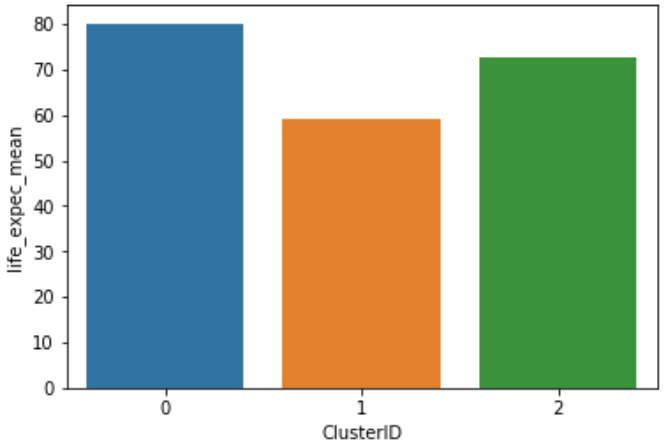
Cluster ID	Cluster Name
0	Developed countries
1	Underdeveloped countries
2	Developing countries

Total fertility rate



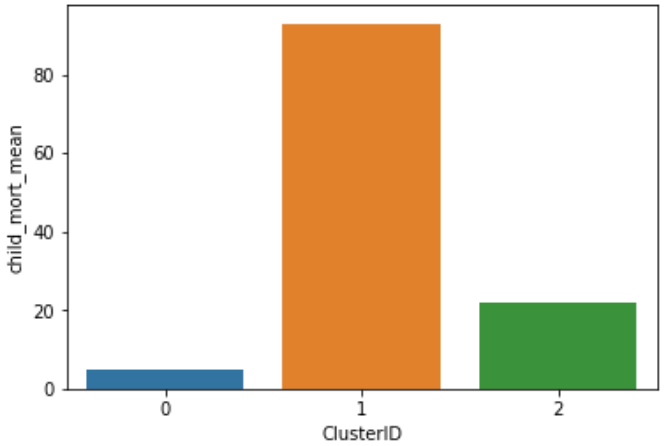
The underdeveloped countries have the highest child fertility rate hence are subject to overpopulation

Life Expectancy



The underdeveloped countries have the lowest life expectancy which is below 60 years of age

Child mortality Rate

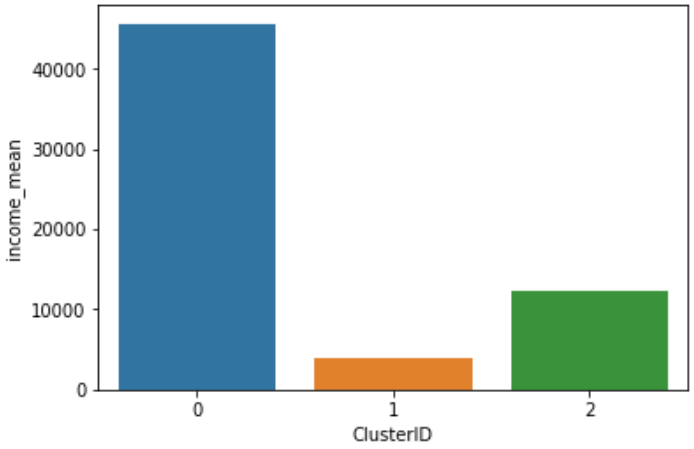


The underdeveloped countries have the highest child mortality rate. Hence life expectancy is low

K-MEANS CLUSTERING - RESULTS

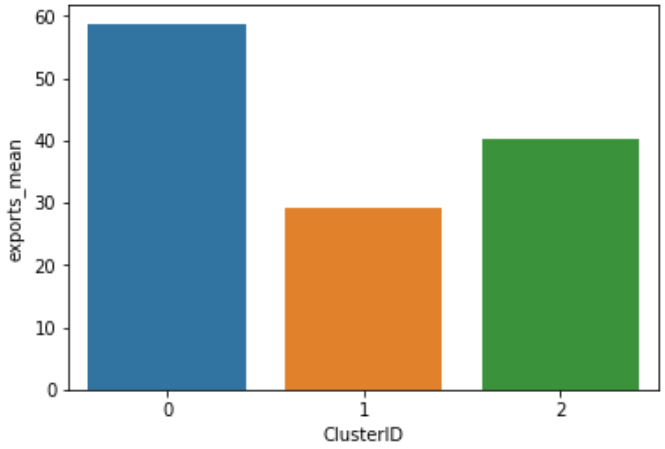
Cluster ID	Cluster Name
0	Developed countries
1	Underdeveloped countries
2	Developing countries

Income



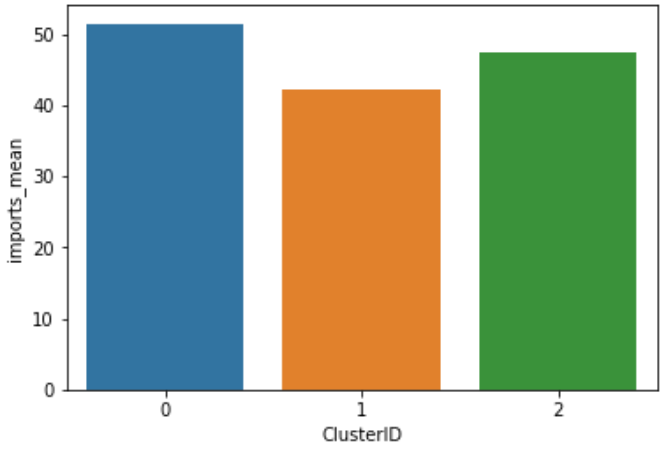
Income is lowest of the underdeveloped countries

Exports



Underdeveloped countries export their products the lowest

Imports

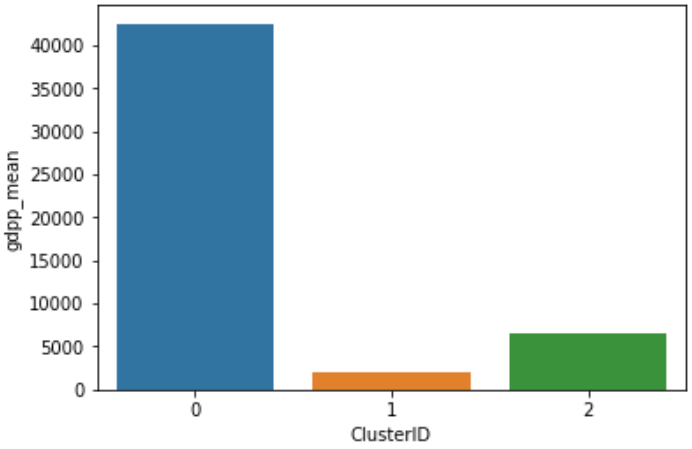


Imports are also lowest of the underdeveloped countries

K-MEANS CLUSTERING - RESULTS

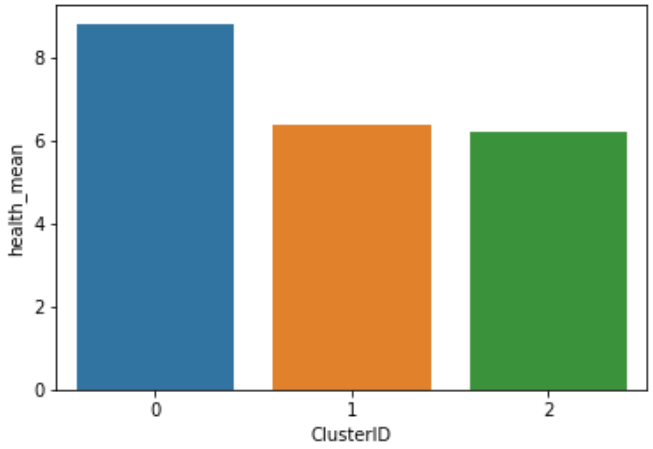
Cluster ID	Cluster Name
0	Developed countries
1	Under-developed countries
2	Developing countries

GDPP



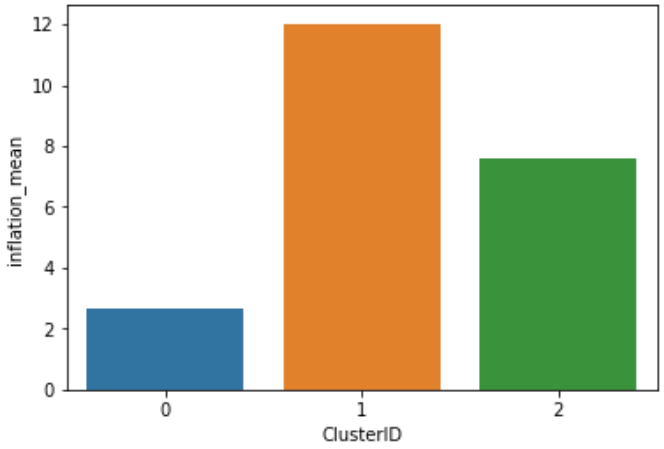
GDPP of the underdeveloped countries is lowest unsurprisingly

Health Spending



Health spending is lowest in developing countries

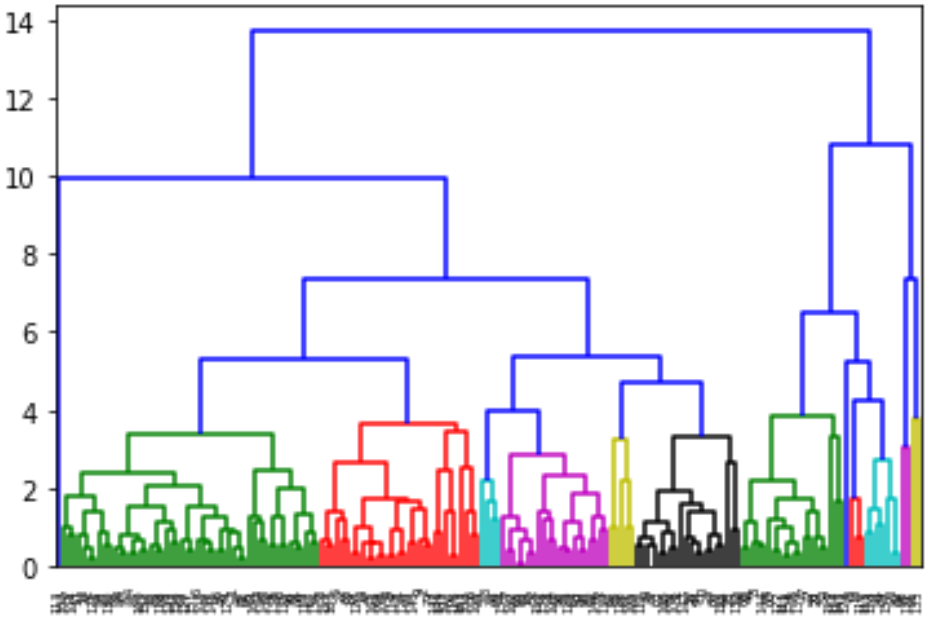
Inflation Rate



Inflation rate is highest in underdeveloped countries

HIERARCHAL CLUSTERING

Dendrogram results



After making a dendrogram using complete method, we selected 2 clusters

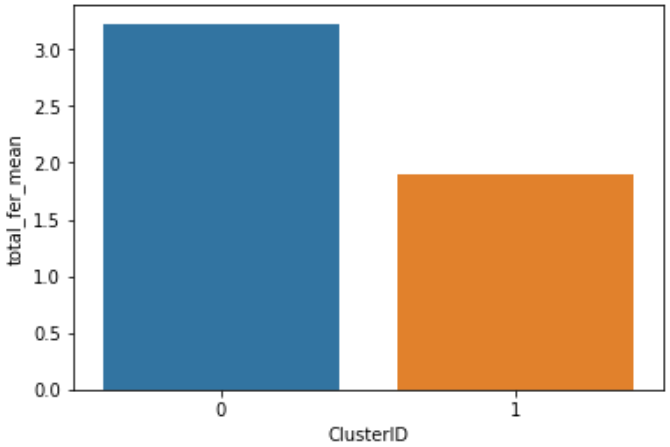
ClusterID	country
0	132
1	35

- Cluster 0 – Undeveloped countries
- Cluster 1 –Developed countries

HIERARCHAL CLUSTERING - RESULTS

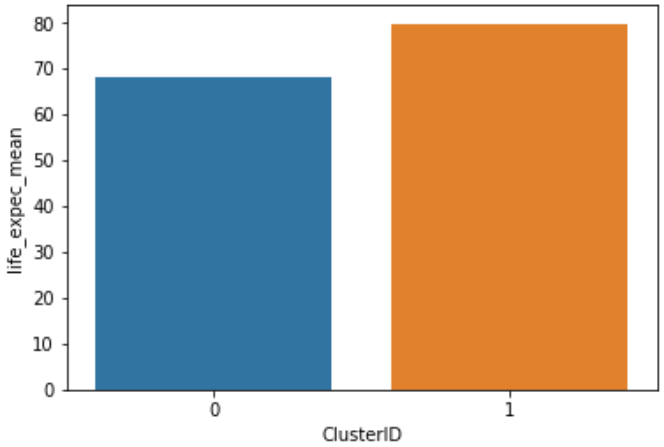
Cluster ID	Cluster Name
0	Undeveloped countries
1	Developed countries

Total fertility rate



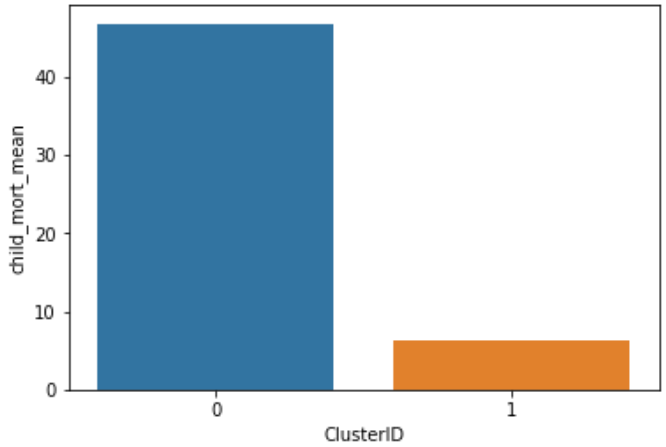
The undeveloped countries have the highest child fertility rate hence are subject to overpopulation

Life Expectancy



The undeveloped countries have the lower life expectancy which is below 70 years of age

Child mortality Rate

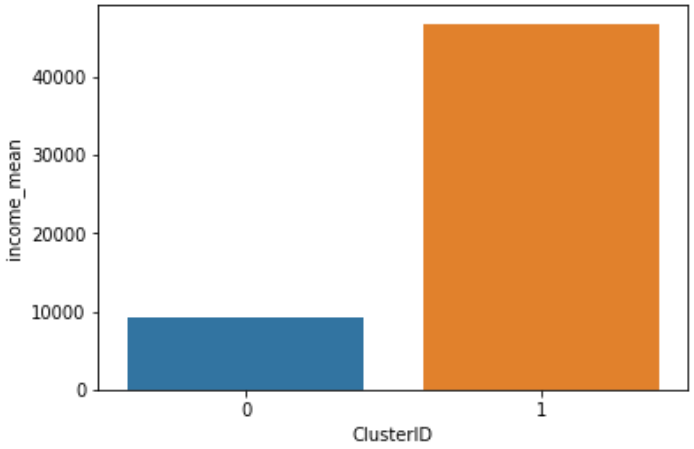


The undeveloped countries have a high child mortality rate. Hence life expectancy is lower

HIERARCHAL CLUSTERING - RESULTS

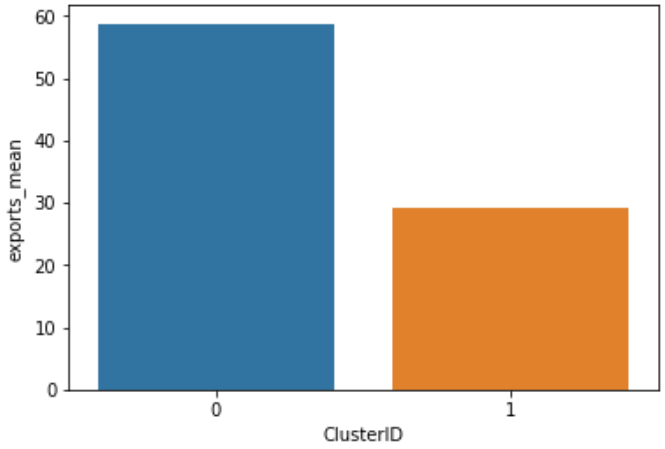
Cluster ID	Cluster Name
0	Undeveloped countries
1	Developed countries

Income



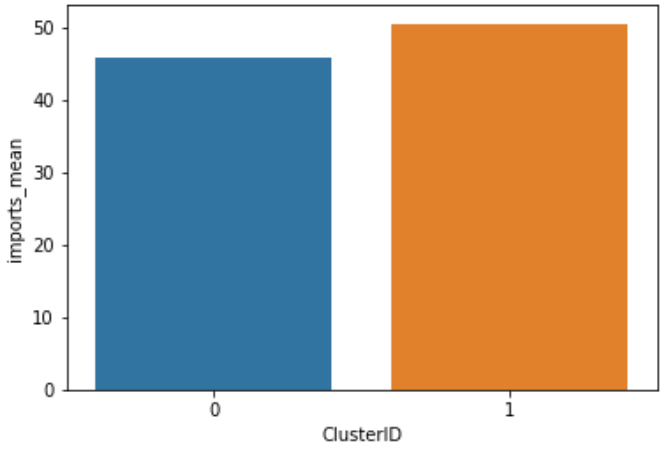
Income is very low of the undeveloped countries

Exports



Exports are lower of the developed countries

Imports

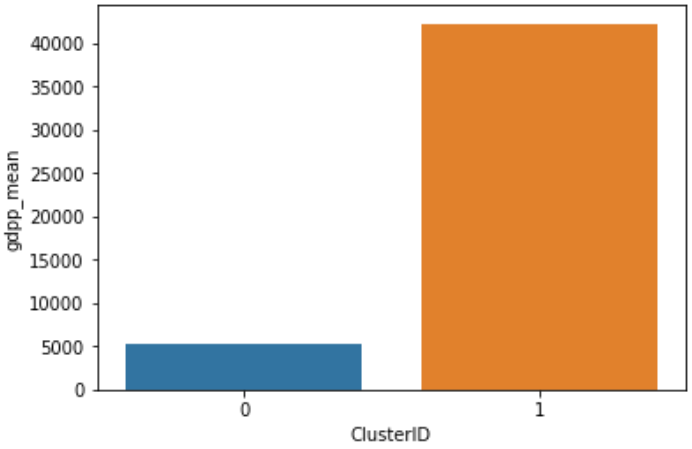


Imports are lower of the undeveloped countries

HIERARCHAL CLUSTERING - RESULTS

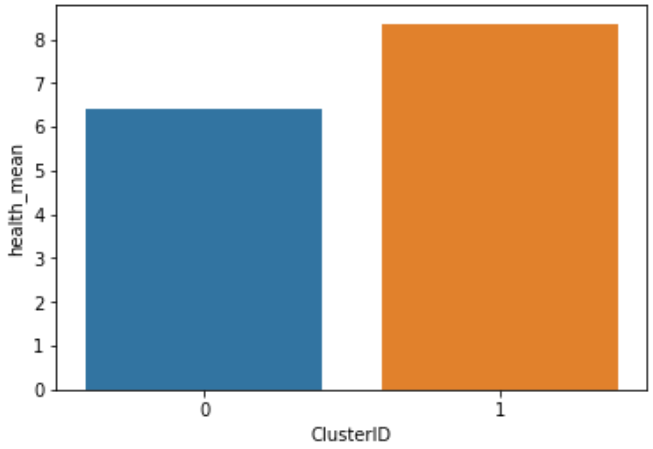
Cluster ID	Cluster Name
0	Developed countries
1	Under-developed countries
2	Developing countries

GDPP



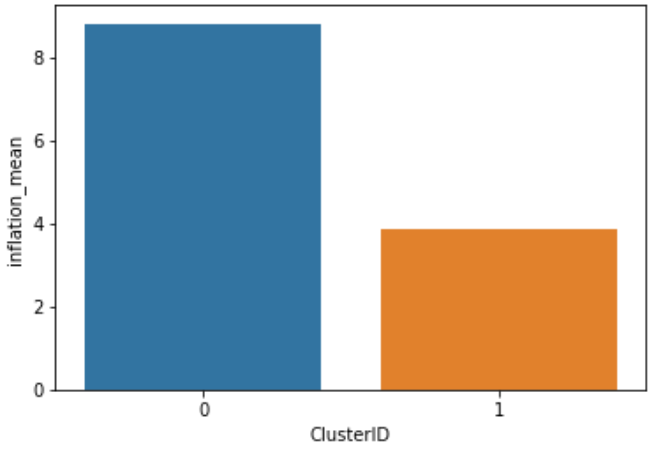
GDPP of undeveloped countries is very low

Health Spending



Health spending is lower in undeveloped countries

Inflation Rate



Inflation rate is higher in undeveloped countries

SELECTED COUNTRIES

We selected 5 countries in the most dire need of aid by looking at the under-developed cluster in the data and analysing all variables.

They are:-

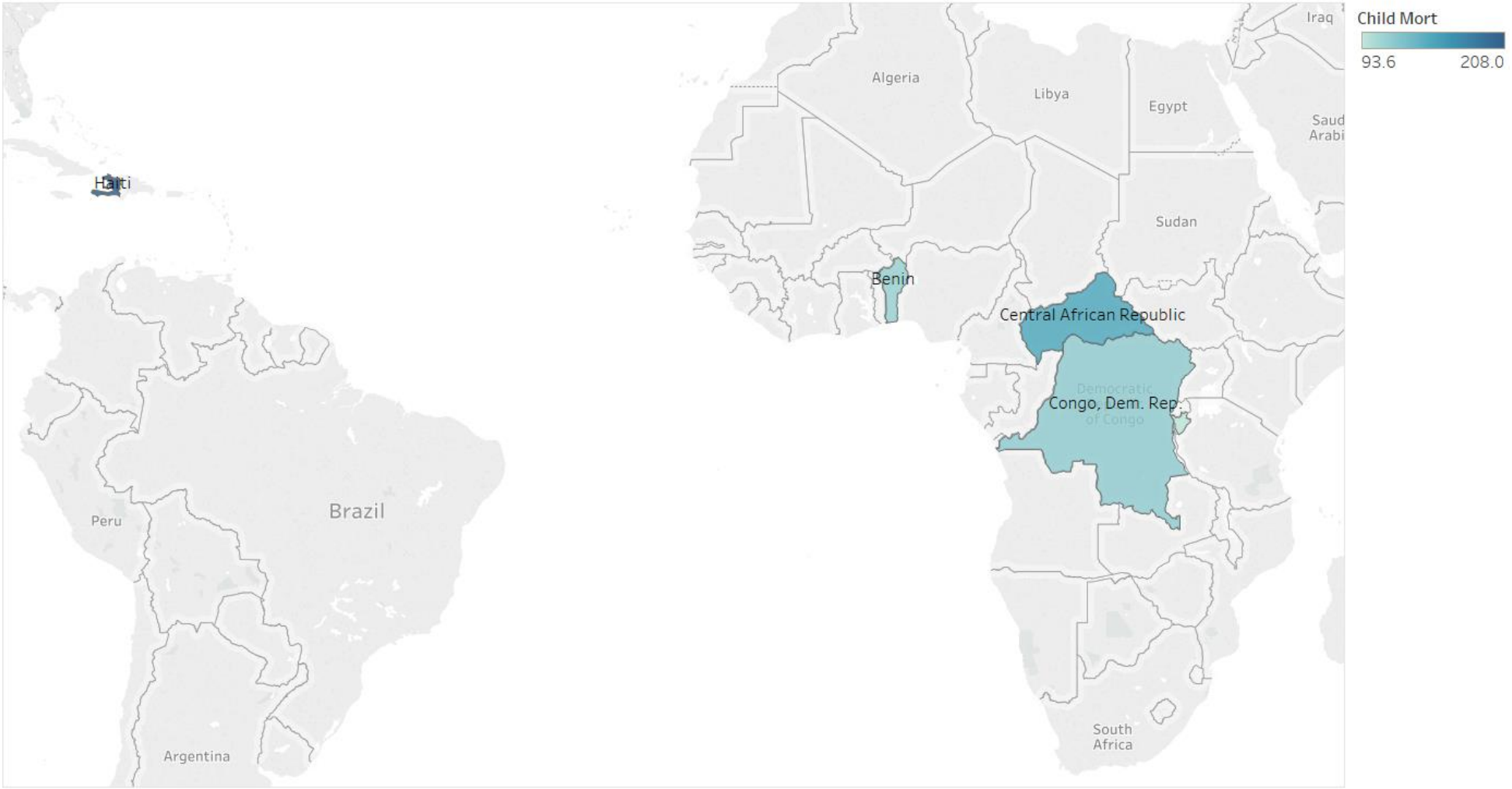
- Benin
- Burundi
- Central African Republic
- Congo, Dem. Rep.
- Haiti

country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	ClusterID
Benin	111	23.8	4.1	37.2	1820	0.885	61.8	5.36	758	1
Burundi	93.6	8.92	11.6	39.2	764	12.3	57.7	6.26	231	1
Central African Republic	149	11.8	3.98	26.5	888	2.01	47.5	5.21	446	1
Congo, Dem. Rep.	116	41.1	7.91	49.6	609	20.8	57.5	6.54	334	1
Haiti	208	15.3	6.91	64.7	1500	5.45	32.1	3.33	662	1

SELECTED COUNTRIES

Comparing the most crucial variables between the selected countries

Child Mortality Rate

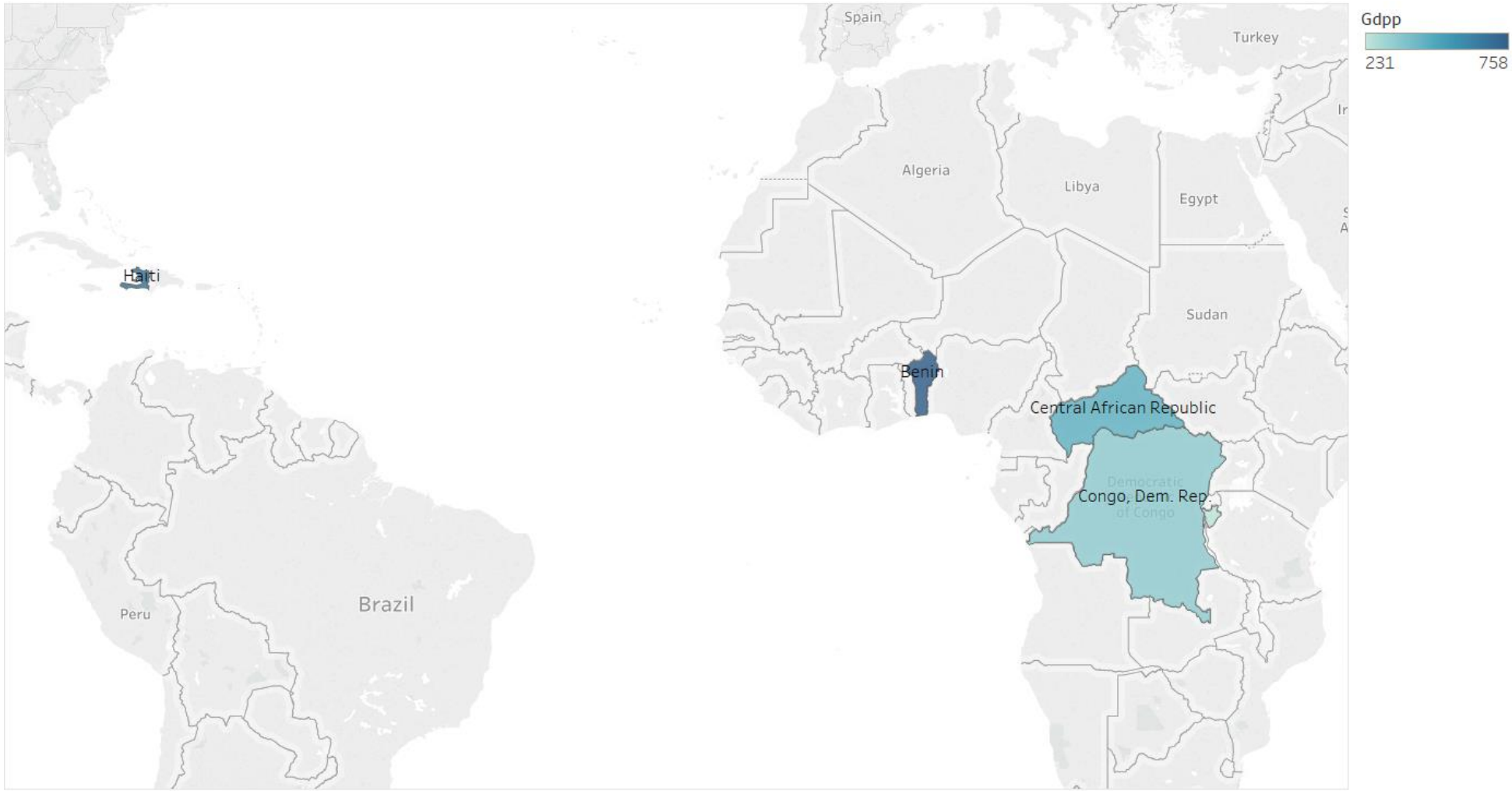


Map based on Longitude (generated) and Latitude (generated). Color shows sum of Child Mort. The marks are labeled by Country. Details are shown for Country.

SELECTED COUNTRIES

Comparing the most crucial variables between the selected countries

GDPP

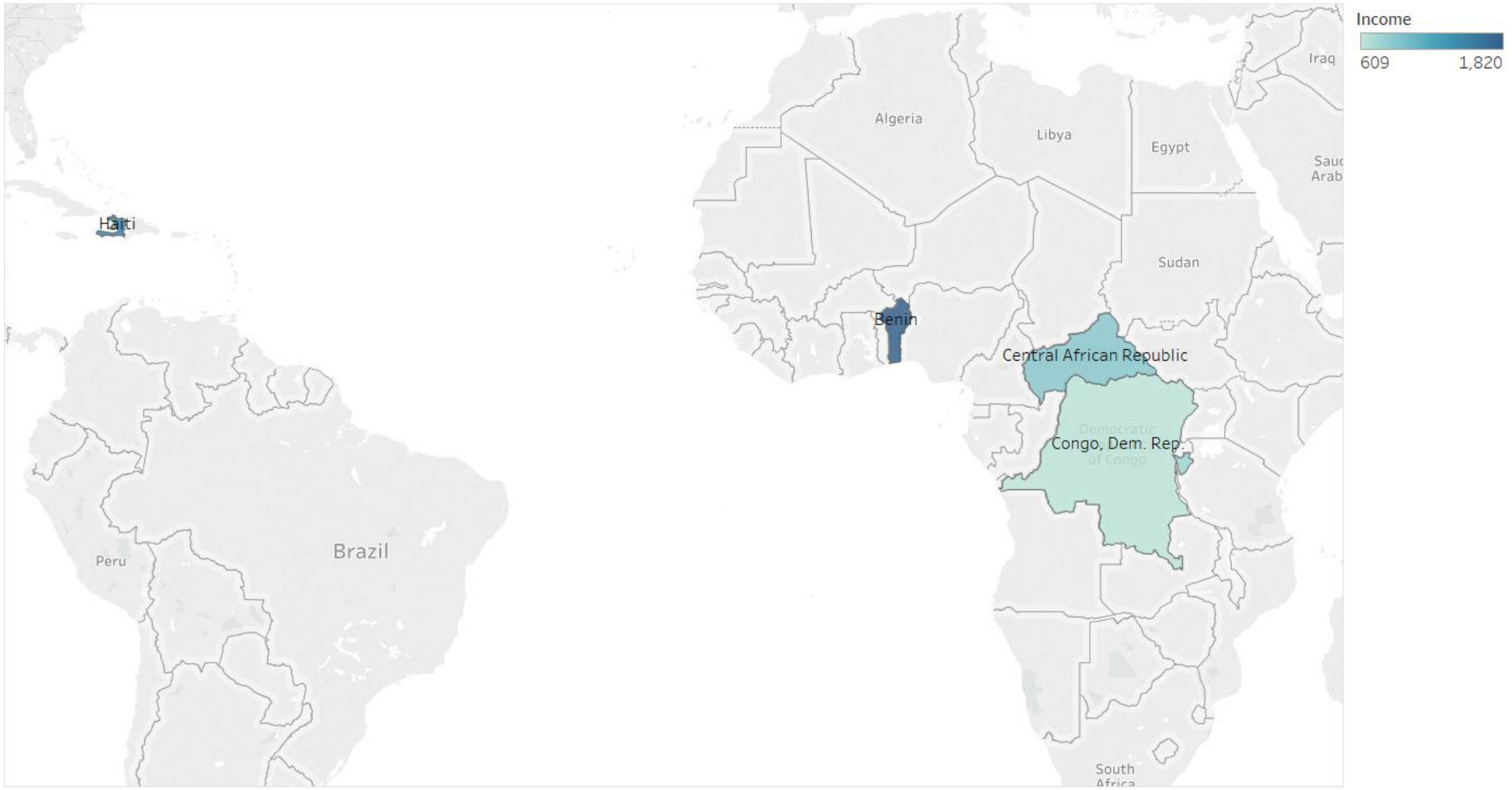


Map based on Longitude (generated) and Latitude (generated). Color shows sum of Gdpp. The marks are labeled by Country. Details are shown for Country.

SELECTED COUNTRIES

Comparing the most crucial variables between the selected countries

Income

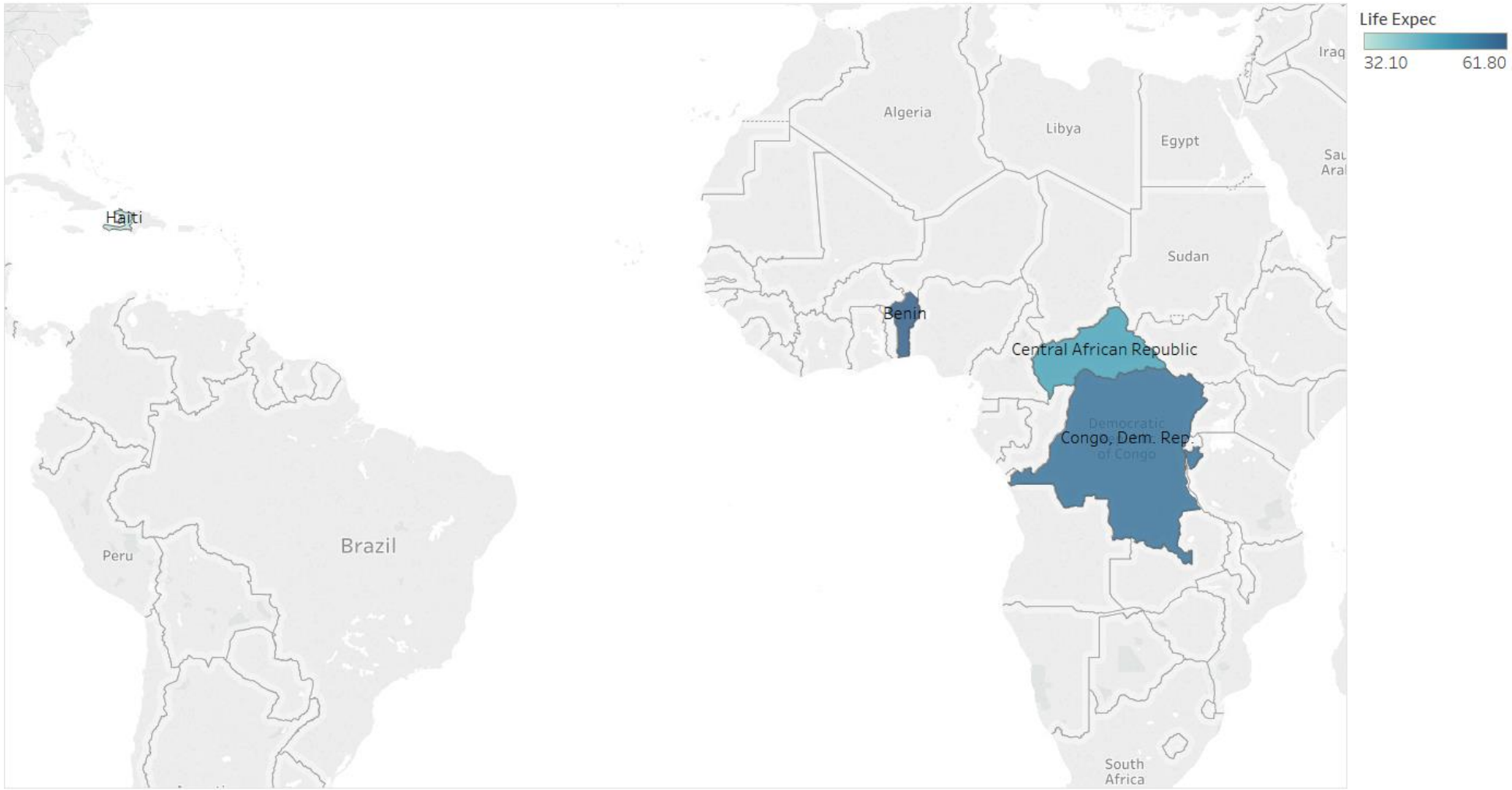


Map based on Longitude (generated) and Latitude (generated). Color shows sum of Income. The marks are labeled by Country. Details are shown for Country.

SELECTED COUNTRIES

Comparing the most crucial variables between the selected countries

Life Expectancy



Map based on Longitude (generated) and Latitude (generated). Color shows sum of Life Expec. The marks are labeled by Country. Details are shown for Country.