Clustering Countries

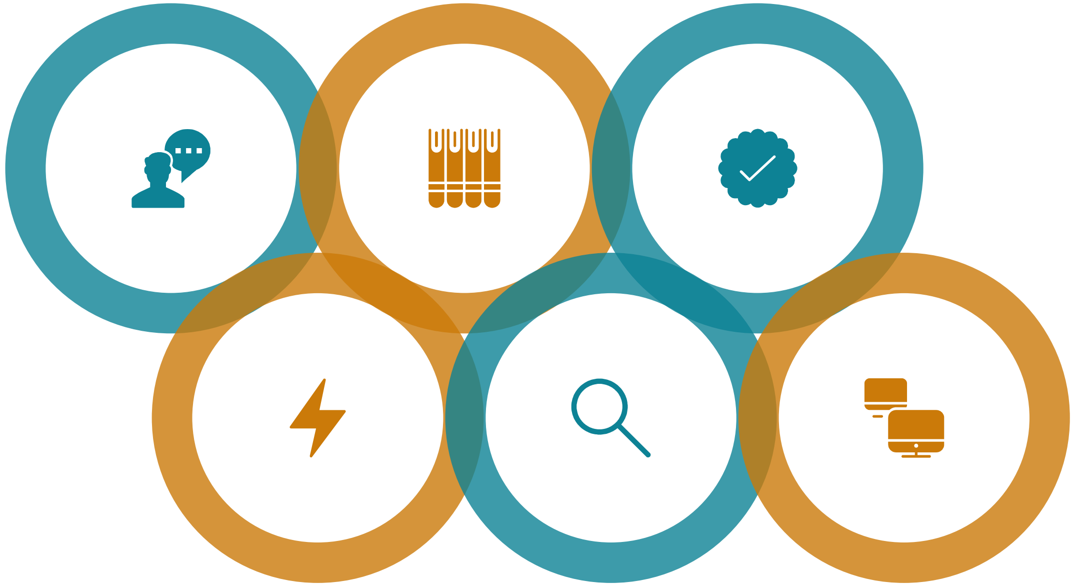
Assignment

By

Sai Phani Rajasekhar Chennapragada

Problem Statement

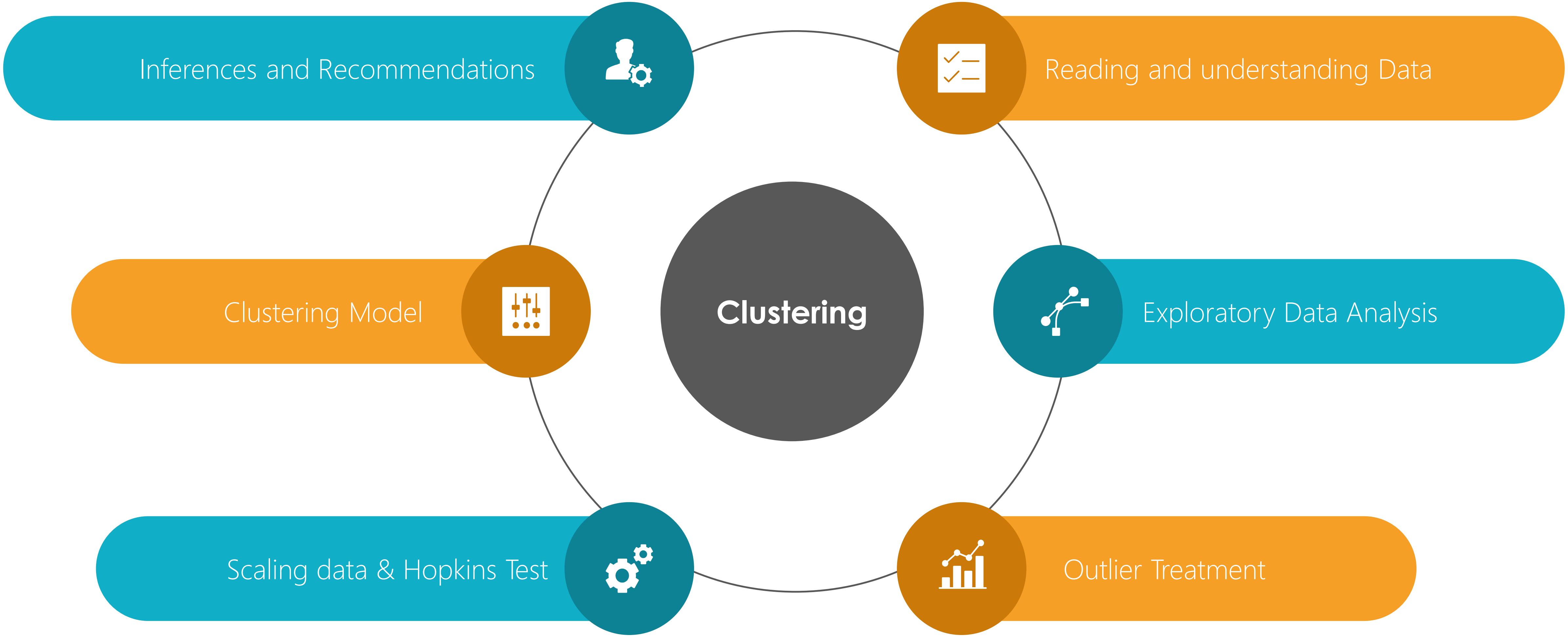
HELP International is an international humanitarian NGO that is committed to fighting poverty and providing the people of backward countries with basic amenities and relief during the time of disasters and natural calamities and they have been able to raise around 10 million dollars. Now the CEO of the NGO needs to decide how to use this money strategically and effectively on backward countries. **Our main task is to cluster the countries by the factors mentioned above and then present solution and recommendations to the CEO**



I've used **K-means** and **Hierarchical clustering** methods to create the **clusters** based on the demographics given to us and finally found the top 10 backward countries that are in the direst need of aid.

Steps for Clustering

The following steps are used in clustering to find the top 10 backward countries in the process of clustering:



Clustering

Reading and understanding Data

Exploratory Data Analysis

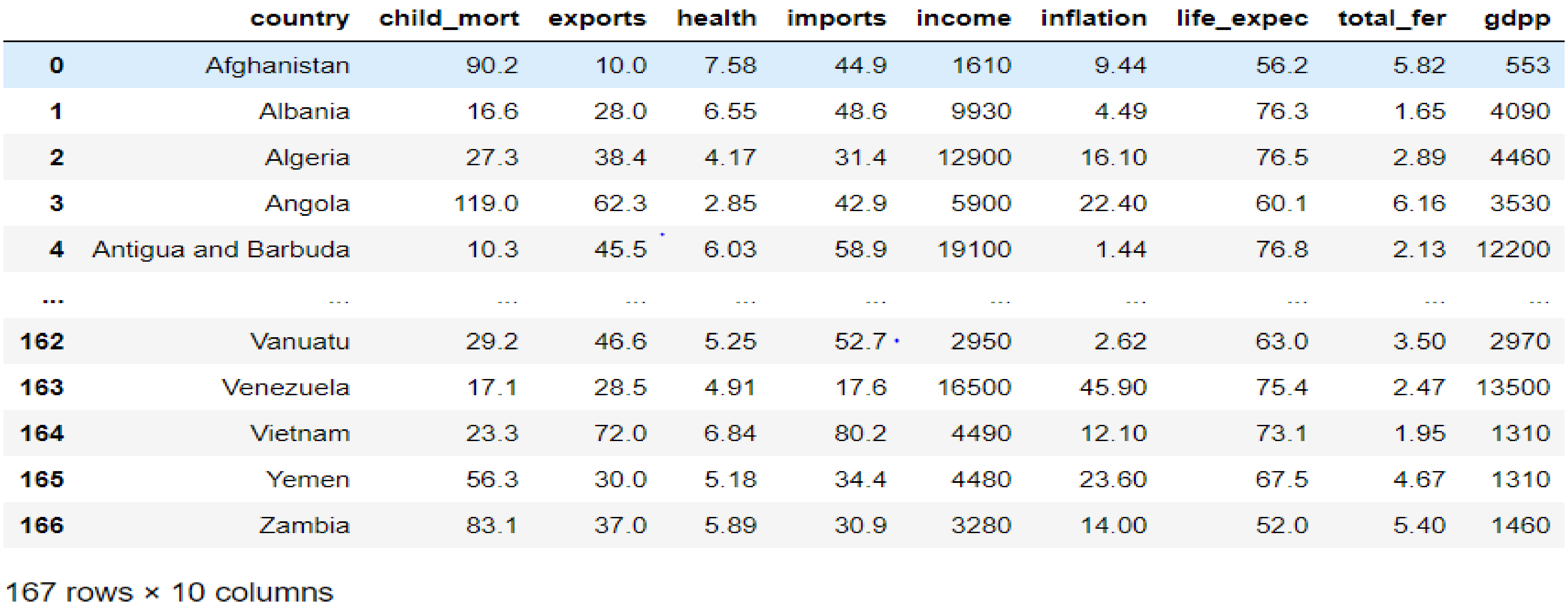
Outlier Treatment

Scaling data & Hopkins Test

Clustering Model

Inferences and Recommendations

Reading &understanding Data

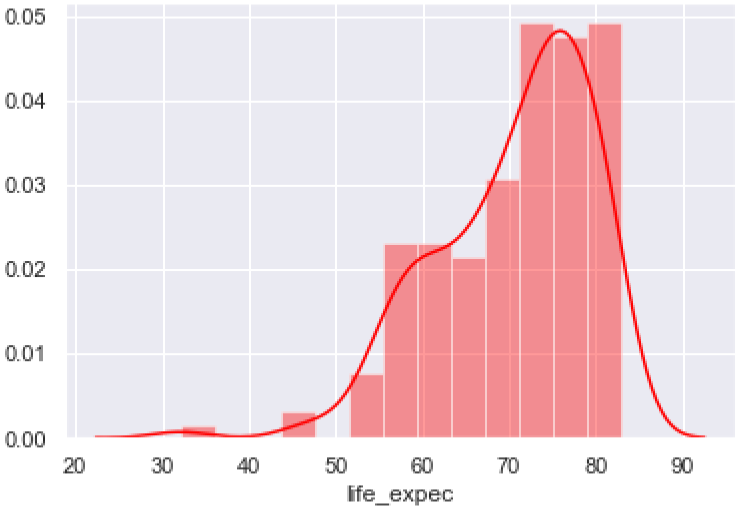


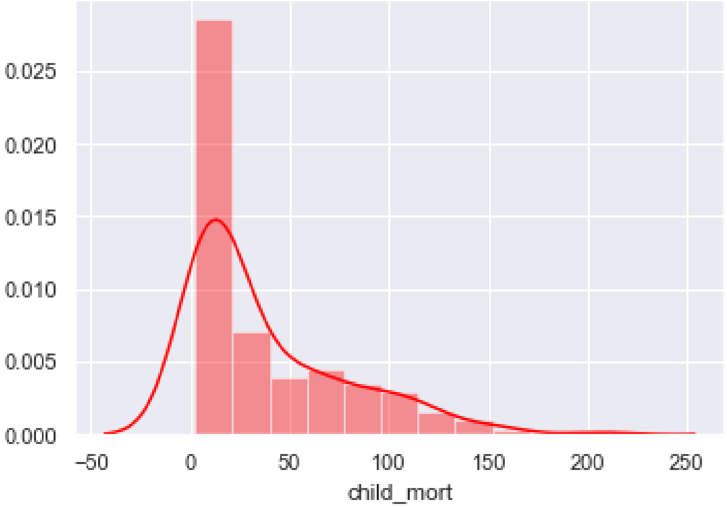
	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
0	Afghanistan	90.2	10.0	7.58	44.9	1610	9.44	56.2	5.82	553
1	Albania	16.6	28.0	6.55	48.6	9930	4.49	76.3	1.65	4090
2	Algeria	27.3	38.4	4.17	31.4	12900	16.10	76.5	2.89	4460
3	Angola	119.0	62.3	2.85	42.9	5900	22.40	60.1	6.16	3530
4	Antigua and Barbuda	10.3	45.5	6.03	58.9	19100	1.44	76.8	2.13	12200
...	...	...	...	...	...	...	...	...	...	...
162	Vanuatu	29.2	46.6	5.25	52.7	2950	2.62	63.0	3.50	2970
163	Venezuela	17.1	28.5	4.91	17.6	16500	45.90	75.4	2.47	13500
164	Vietnam	23.3	72.0	6.84	80.2	4490	12.10	73.1	1.95	1310
165	Yemen	56.3	30.0	5.18	34.4	4480	23.60	67.5	4.67	1310
166	Zambia	83.1	37.0	5.89	30.9	3280	14.00	52.0	5.40	1460

167 rows × 10 columns

- As we can see clearly, there are totally 167 countries in the data
- Exports, health and imports are mentioned as the percentage of GDP per capita. So, we have converted these column values to absolute values instead of percentage for better comparison.

Exploratory Data Analysis



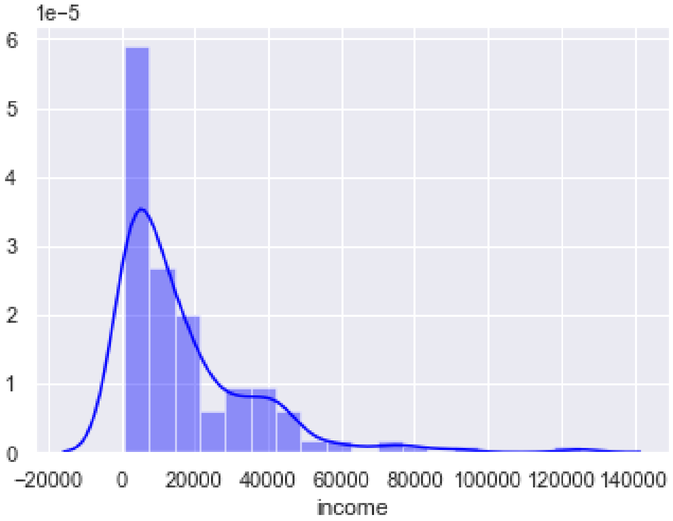


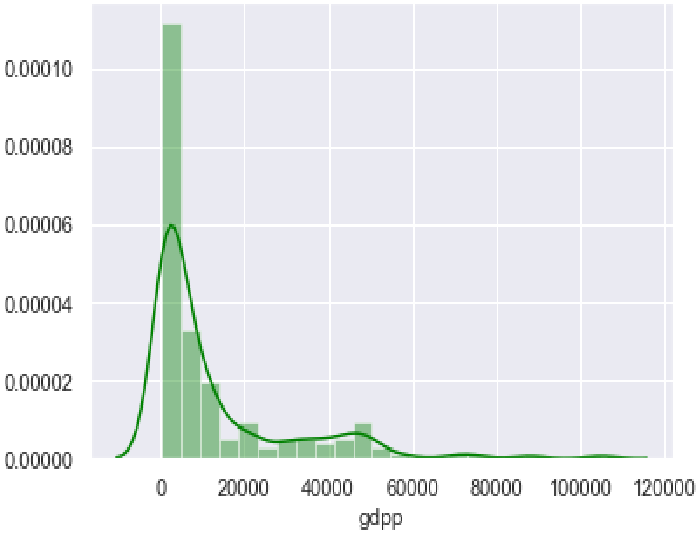
Univariate analysis:

- From the plots, we can see most of the average income per person and GDP per capita for countries are observed in the range of 0-15000

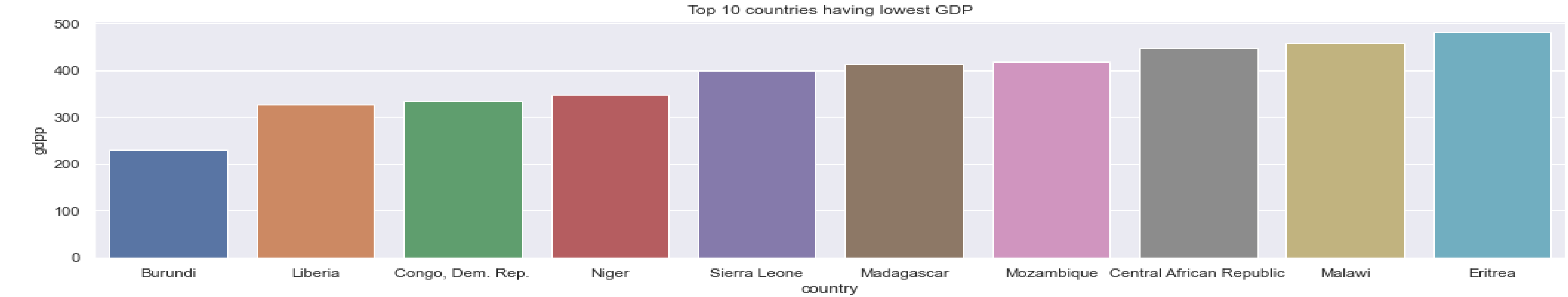
- On average, we can infer that life expectancy of a person for most of the countries is observed between 60-80

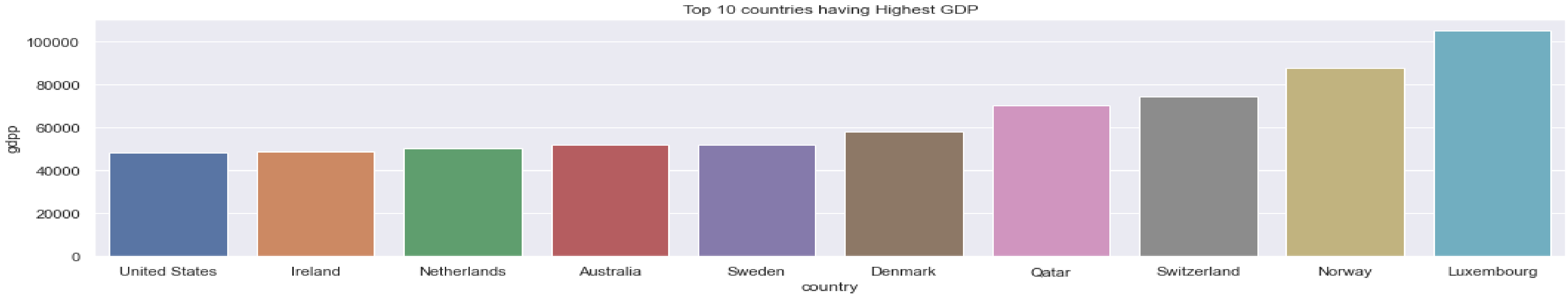
- Child mortality (Death of children under 5 years of age per 1000 live births) seems to be below 50 in most of the countries, only few countries child mortality is above 100.





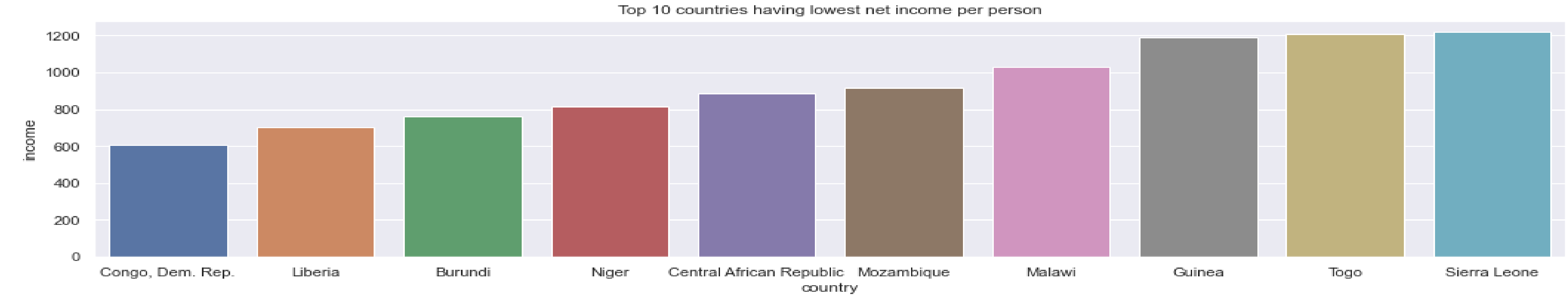
Exploratory Data Analysis

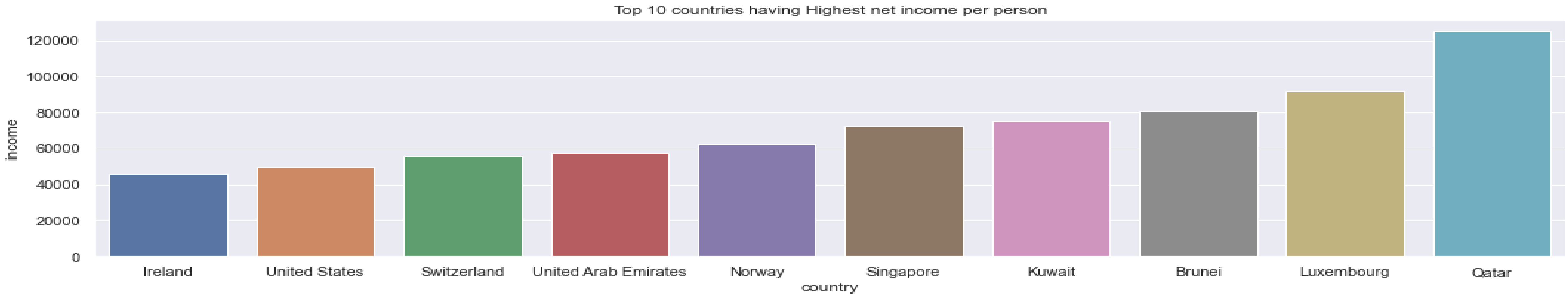

Top 10 countries having lowest GDP


Top 10 countries having Highest GDP

- From the above plot, we can see that **top 5 countries** having lowest GDP are **Burundi, Liberia, Congo, Dem. Rep., Niger and Sierra Leone.**
- Top 5 countries having highest GDP are **Luxembourg, Norway, Switzerland, Qatar and Denmark.**

Exploratory Data Analysis

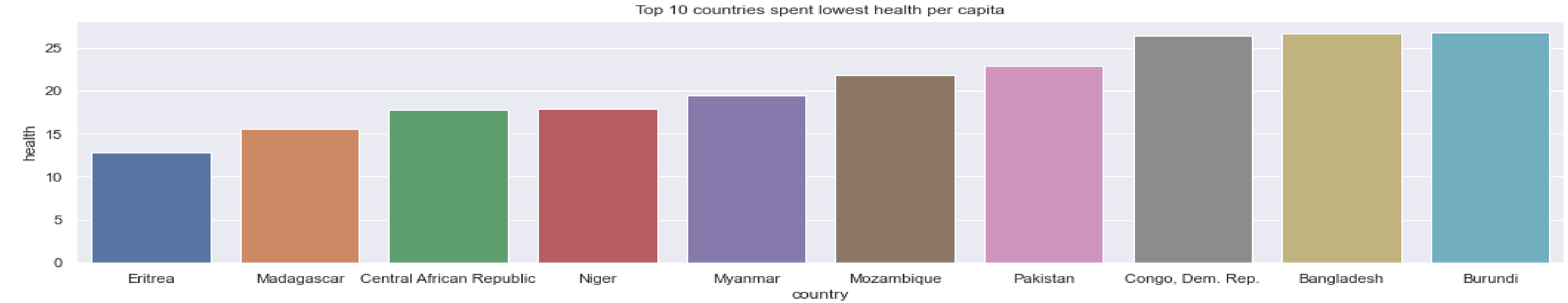


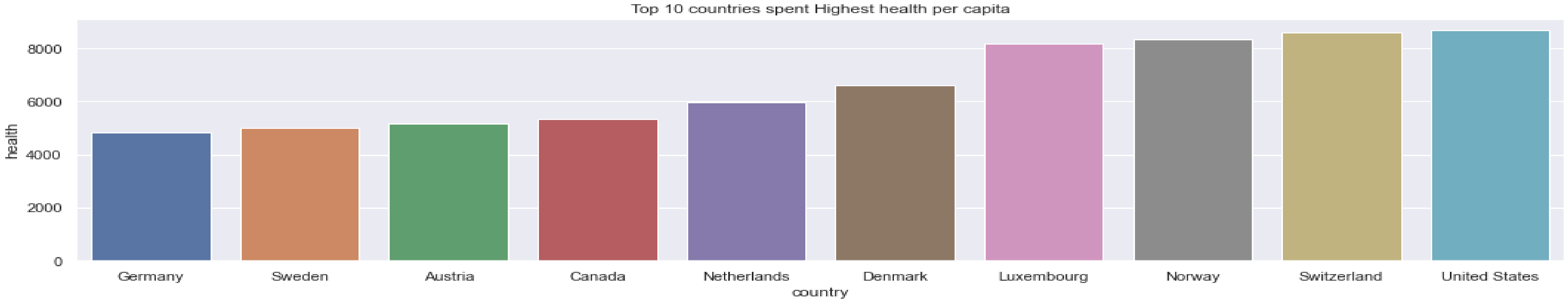


- From the above plot, we can see that **top 5 countries** having lowest net income per person are **Congo, Dem. Rep., Liberia, Burundi, Niger and Central African Republic.**

- **Top 5 countries** having highest net income per person are **Qatar, Luxembourg, Brunei, Kuwait and Singapore.**

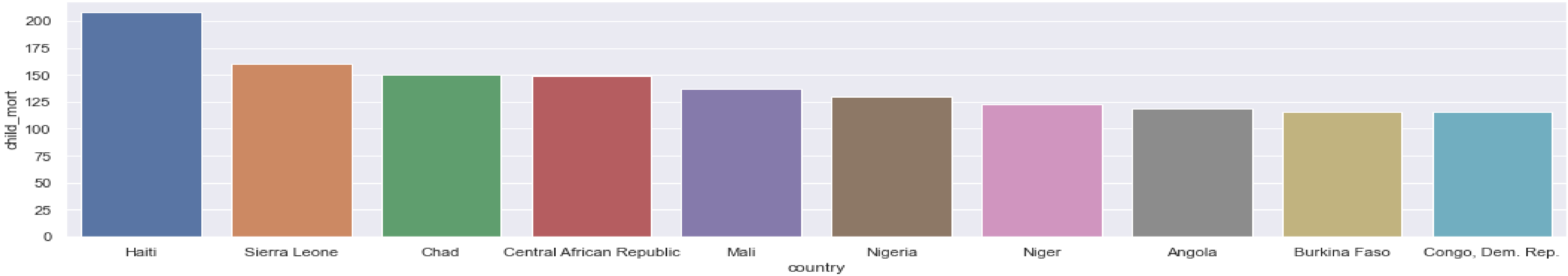
Exploratory Data Analysis



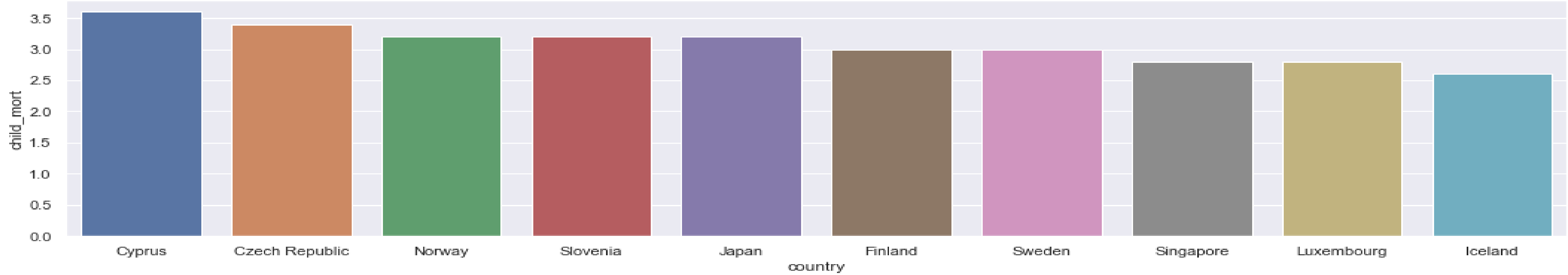


- From the above plot, we can see that **top 5 countries spent lowest health per capita** are **Eritrea, Madagascar, Central African Republic, Niger and Myanmar**
- **Top 5 countries spent highest health per capita** are **United States, Switzerland, Norway, Luxembourg and Denmark.**

Exploratory Data Analysis



Top 10 countries high Death of children under 5 years of age per 1000 live births

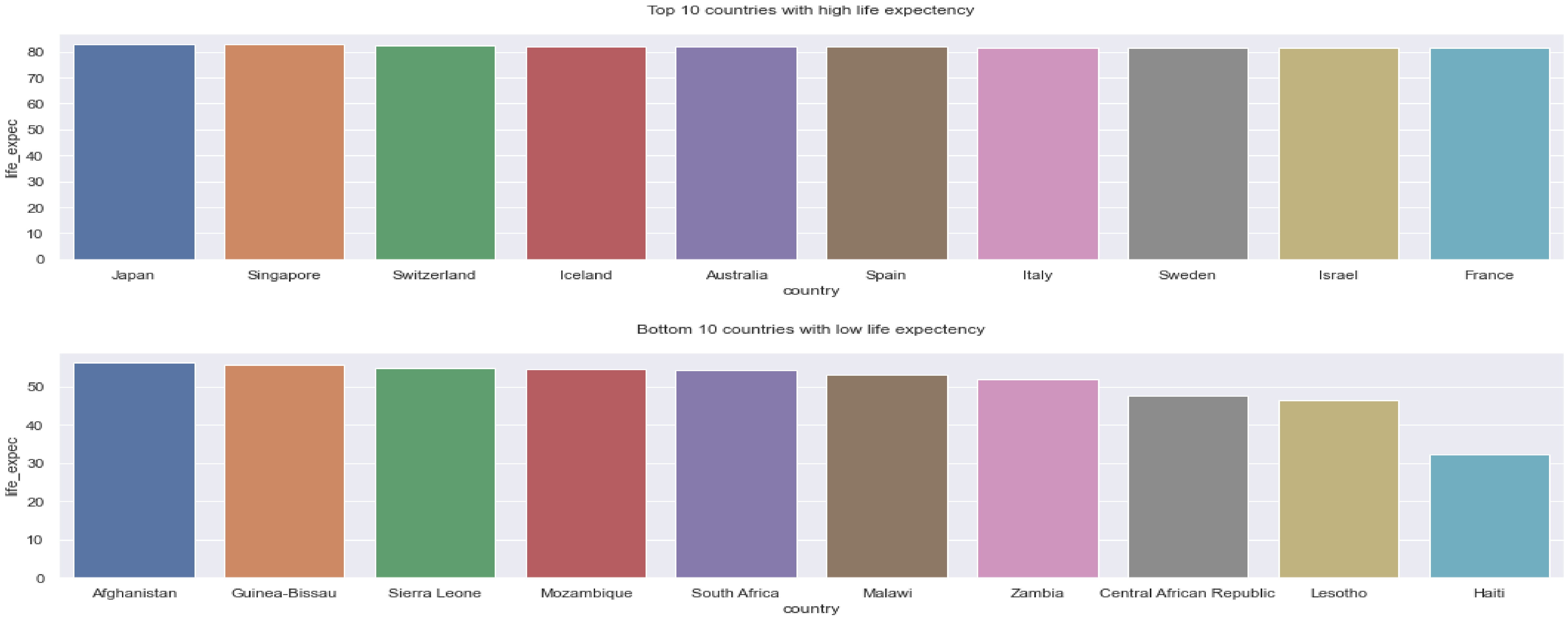


Bottom 10 countries low Death of children under 5 years of age per 1000 live births

- From the above plot, we can see that **top 5 countries** with **lowest child mortality** are **Iceland, Luxembourg, Singapore, Sweden and Finland**

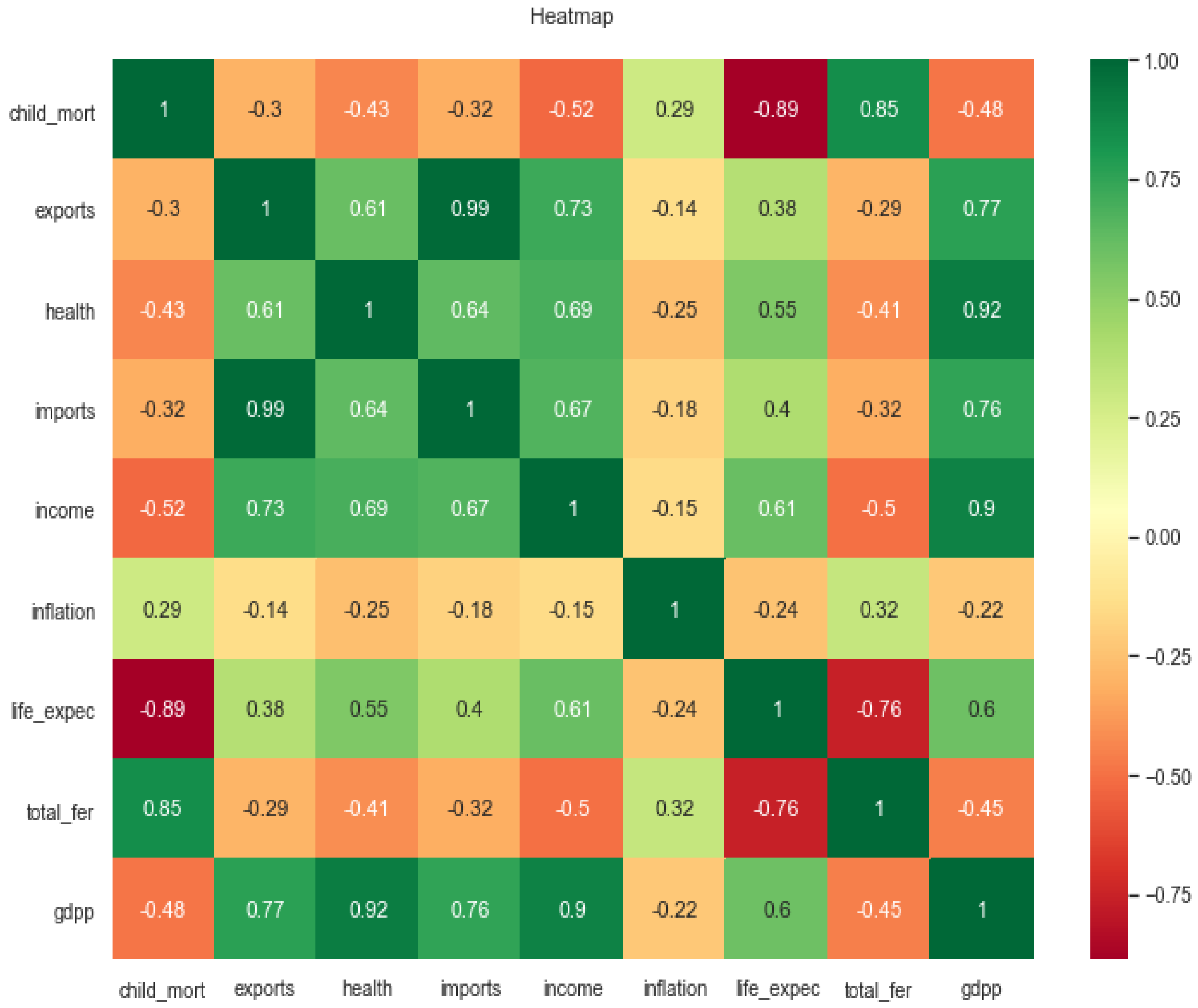
- Top 5 countries with **highest child mortality** are **Haiti, Sierra Leone, Chad, Central African Republic and Mali.**

Exploratory Data Analysis



- From the above plot, we can see that top 10 countries have the same life expectancy of **80-82 years approximately**.
- Top 5 countries with **lowest life expectancy** are **Haiti, Lesotho, Central African Republic, Zambia and Malawi.**

Exploratory Data Analysis



Heatmap

	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
child_mort	1	-0.3	-0.43	-0.32	-0.52	0.29	-0.89	0.85	-0.48
exports	-0.3	1	0.61	0.99	0.73	-0.14	0.38	-0.29	0.77
health	-0.43	0.61	1	0.64	0.69	-0.25	0.55	-0.41	0.92
imports	-0.32	0.99	0.64	1	0.67	-0.18	0.4	-0.32	0.76
income	-0.52	0.73	0.69	0.67	1	-0.15	0.61	-0.5	0.9
inflation	0.29	-0.14	-0.25	-0.18	-0.15	1	-0.24	0.32	-0.22
life_expec	-0.89	0.38	0.55	0.4	0.61	-0.24	1	-0.76	0.6
total_fer	0.85	-0.29	-0.41	-0.32	-0.5	0.32	-0.76	1	-0.45
gdpp	-0.48	0.77	0.92	0.76	0.9	-0.22	0.6	-0.45	1

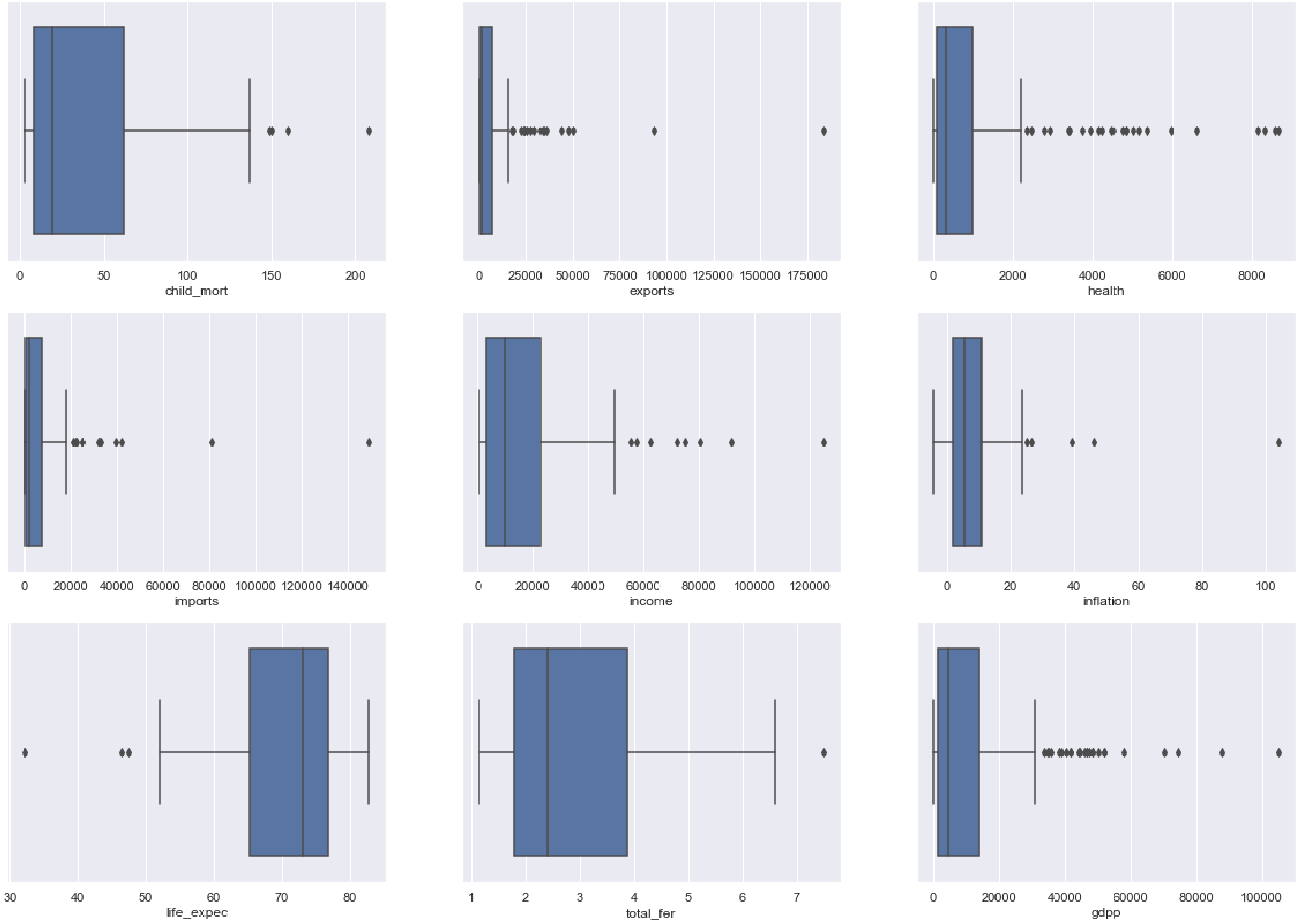
- From the heat map, we can infer that there is a high correlation between health, income and GDP, so, if average net income per person increases the GDP of the country will also increase.
- If child mortality(Death of children under 5 years of age per 1000 live births) decreases the life expectancy will also decreases.
- If you spend more on Health , then the child mortality will reduce and it will increase the GDP rate.
- Also we can see that there is high correlation between Imports and exports, GDP and health, GDP and income

Box plot before Outlier Treatment

Outlier Analysis:

Plotting outlier analysis of all the columns in the data and observed there are outliers

- The main objective is to we need to keep the low range outliers for countries with low child mortality ,low inflation, low total fertility.
- We need to remove the high range outliers for countries with high exports ,health , imports, income, GDP, life expectancy.

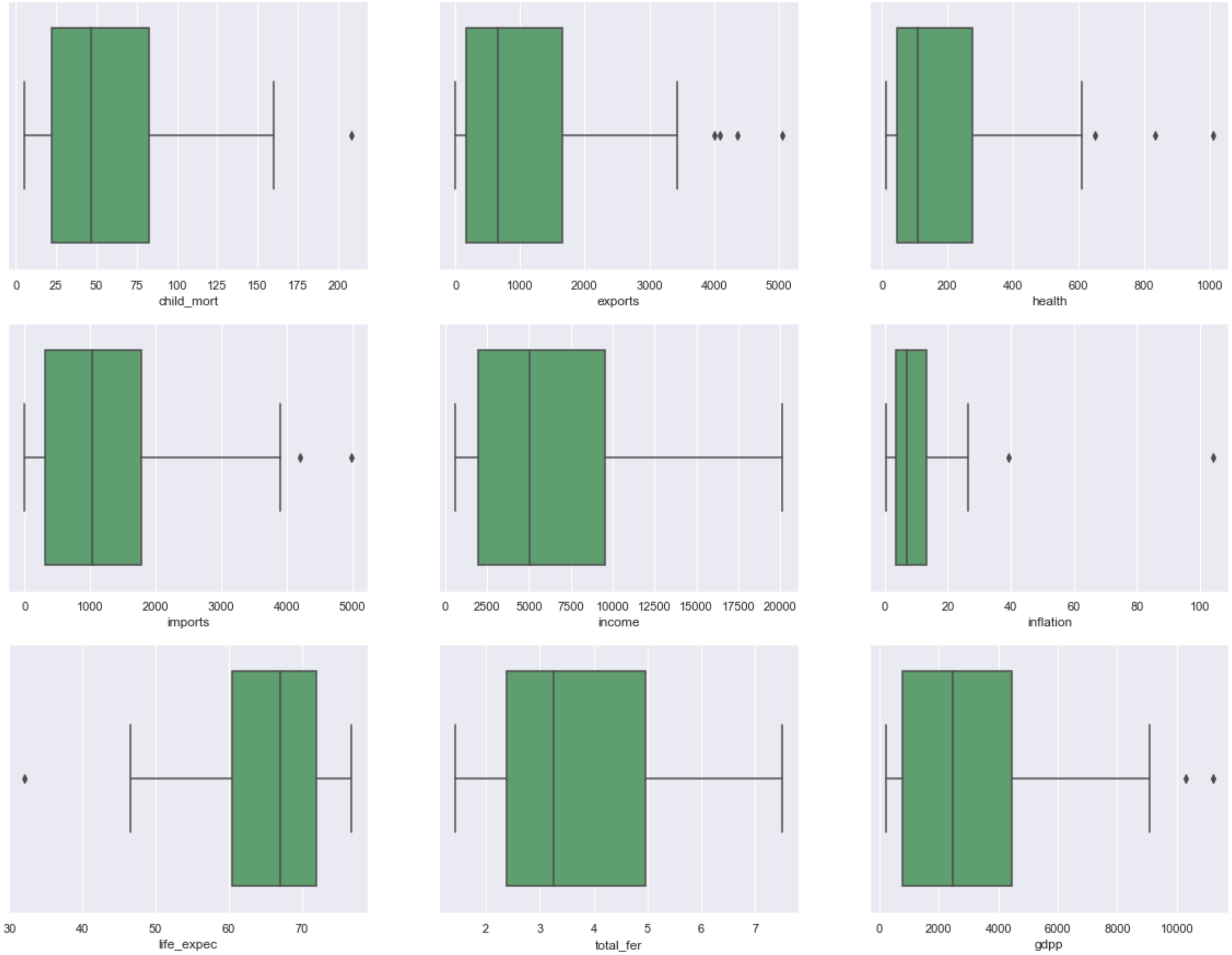


Box plot After Outlier Treatment

Outlier Analysis:

Removing outliers and plotting box plots after removal Listed the steps.

- Removing the low range outliers for countries with low child mortality ,low inflation, low total fertility.
- Removing the high range outliers for countries with high exports , health, imports, income, GDP, life expectancy.
- So followed the method hard capping by removing these outliers with quantiles of 0.05 for low range and 0.95 for high range.



Scaling Data and Hopkins Test

Scaling Data:

- We need to drop the country column from the data and scale the rest of the numerical data.
- Since the data was not in the same units and each column has different range of values, we need to scale the data as it might effect the formation of clusters
- I have used Standard Scaler which converts the data with mean 0 and standard deviation as 1

Hopkins Test:

- 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 {0.7, ..., 0.99}, it has a high tendency to cluster.
- Here we got average value of 0.83 which has high tendency to from clusters with the given data.

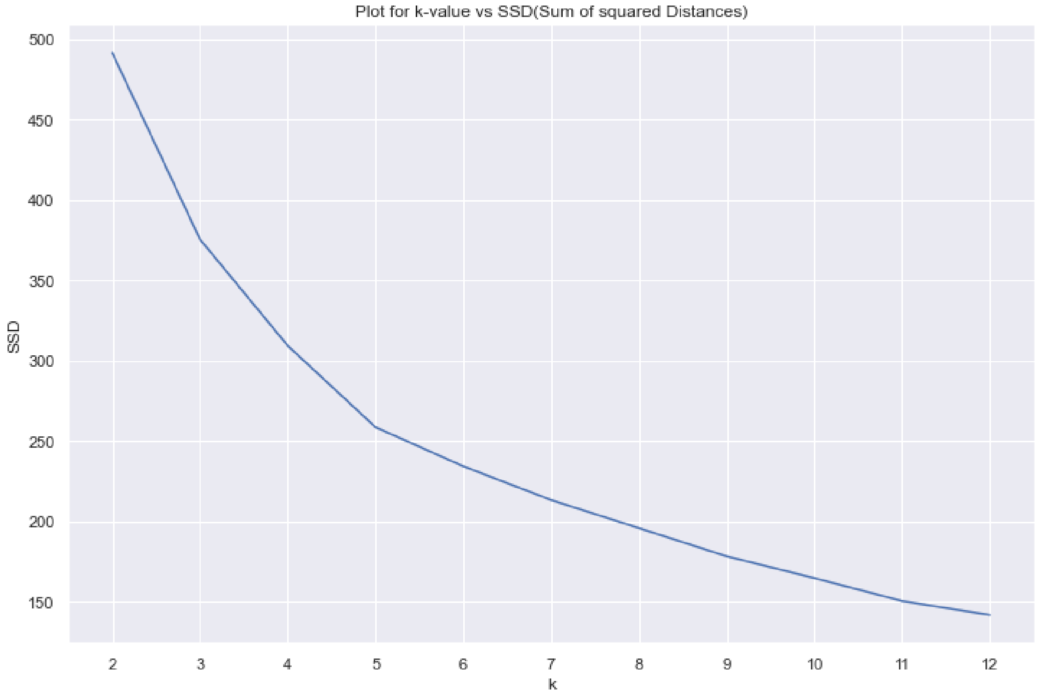
Clustering Model

K-means clustering:

- In k-means clustering, we need to check the value of k (number of clusters) before clustering.
- To find the value of k, we mainly have 2 methods
  - Elbow curve method
  - Silhouette Analysis

Elbow curve method:

- Elbow curve method will be plotted with different k-values and Sum of squared distances (SSD)
- The value of k will be determined where the elbow shape has been formed. We see that for k value 3 and 5 seems to be good values



Elbow Curve Method

Clustering Model

Silhouette Analysis:

silhouette score= $(p-q)/\ max(p,q)$

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.

After performing the analysis we got silhouette score as

For n_clusters=2, the silhouette score is 0.40658955454784557

For n_clusters=3, the silhouette score is 0.3446002070129968

For n_clusters=4, the silhouette score is 0.35634127956386136

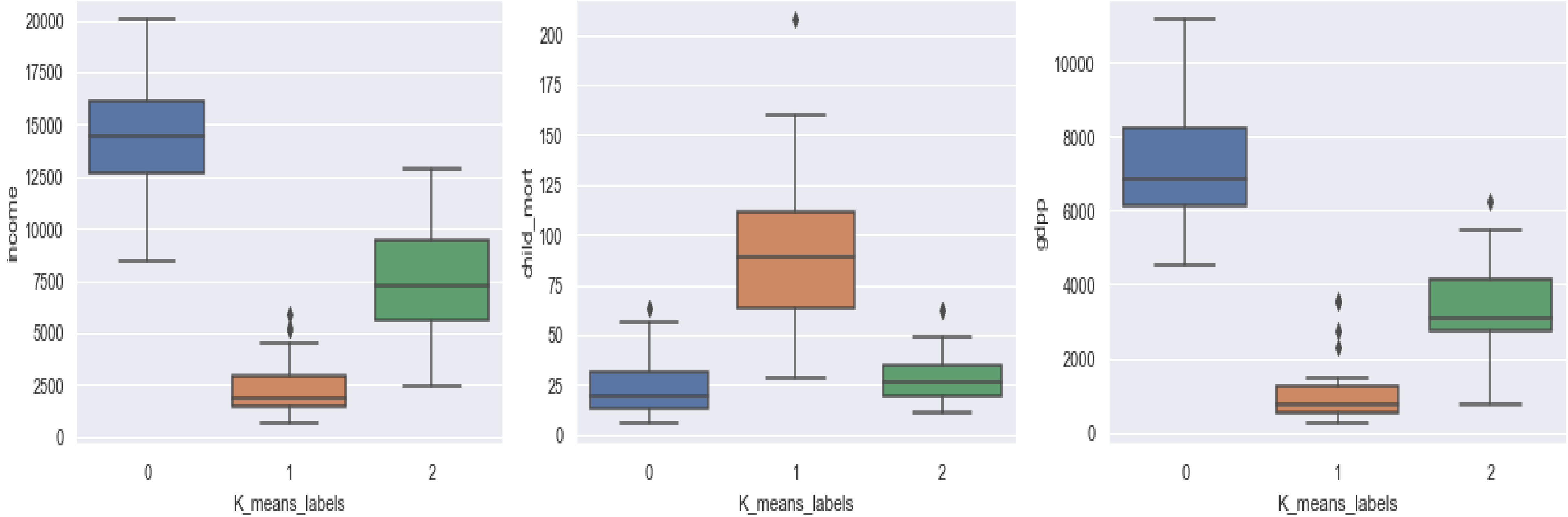
For n_clusters=5, the silhouette score is 0.27369759627280654

For n_clusters=6, the silhouette score is 0.2242479873721014

Lets take the number of clusters as 3 since it seems to be more promising value by elbow curve as well as silhouette score

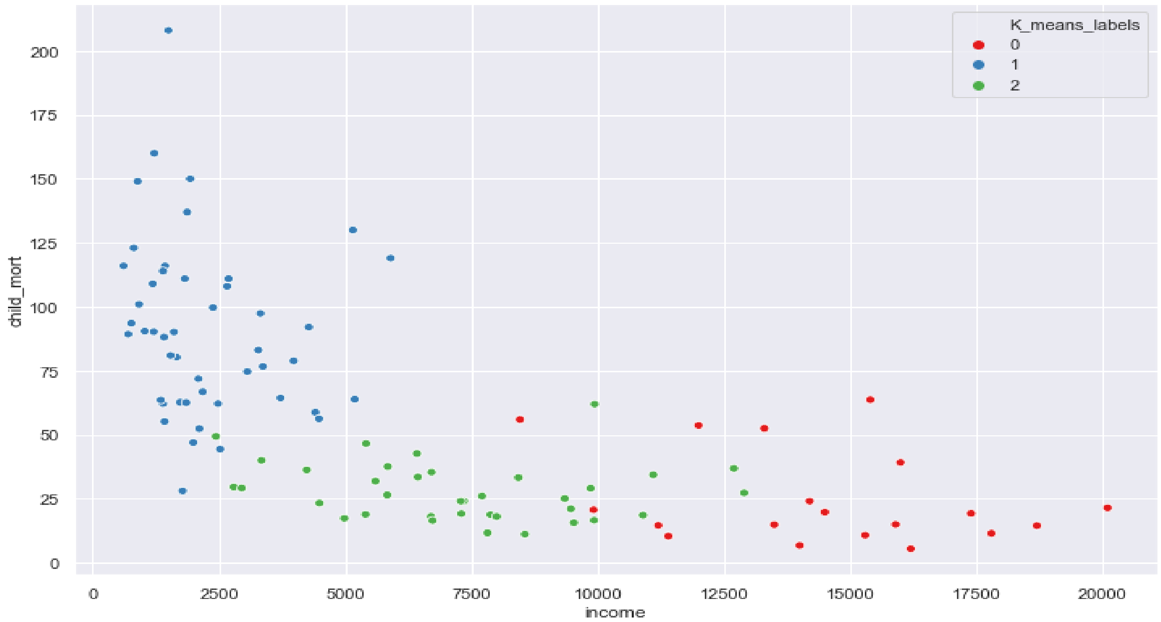
Clustering Model

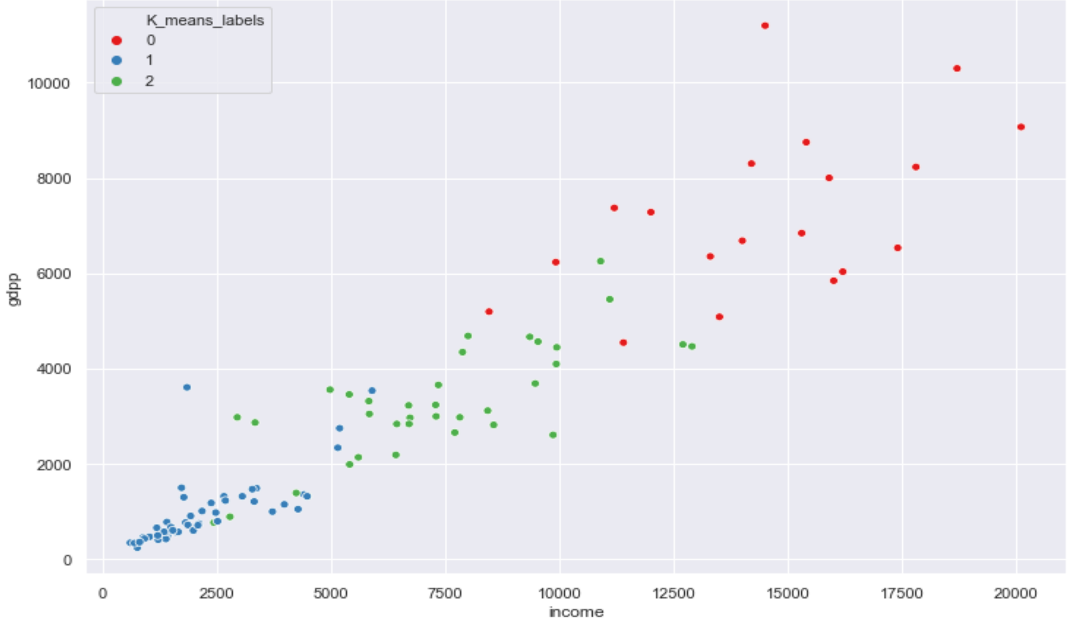
Visualizing the K-means labels formed:



Clustering Model

Visualizing scatter plots for K-means labels formed:



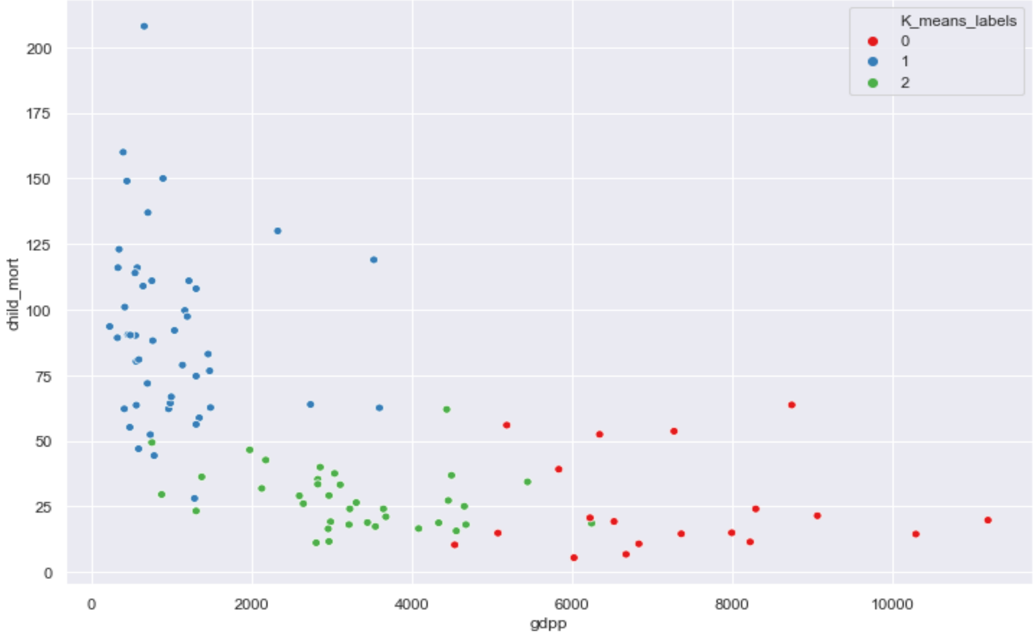


Finally we can see that using, **k-means clustering** method, clusters have been formed as

label = 0, high income, high GDP and low child mortality

label = 1, low income, low GDP and high child mortality

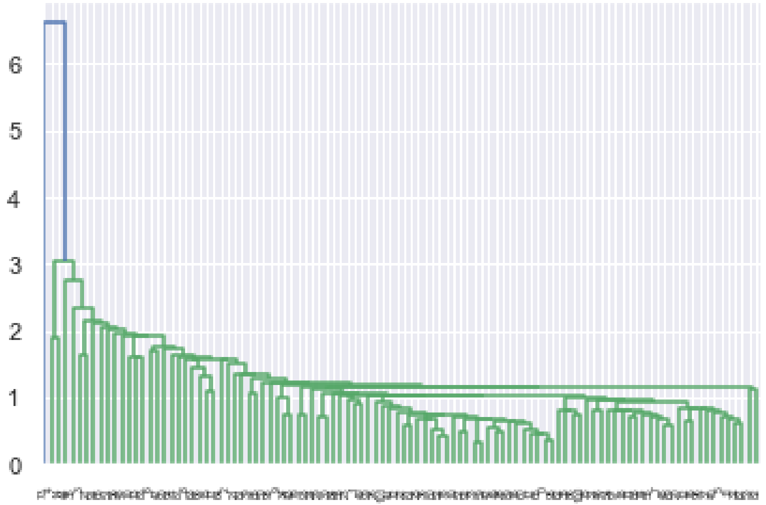
label = 2, moderate income, moderate GDP and low child mortality



Clustering Model

Hierarchical clustering:

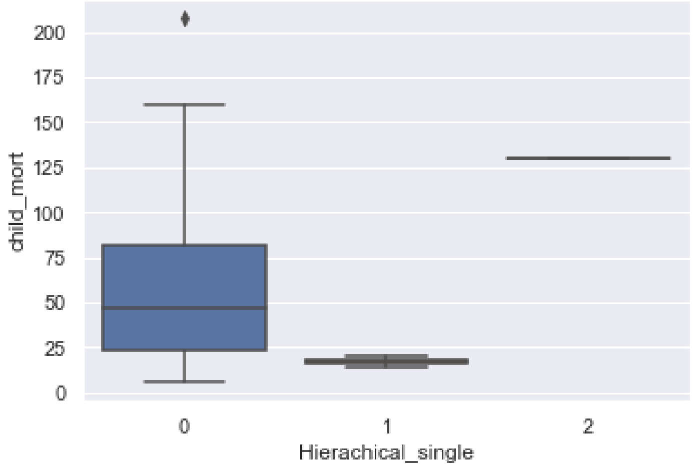
Single linkage:

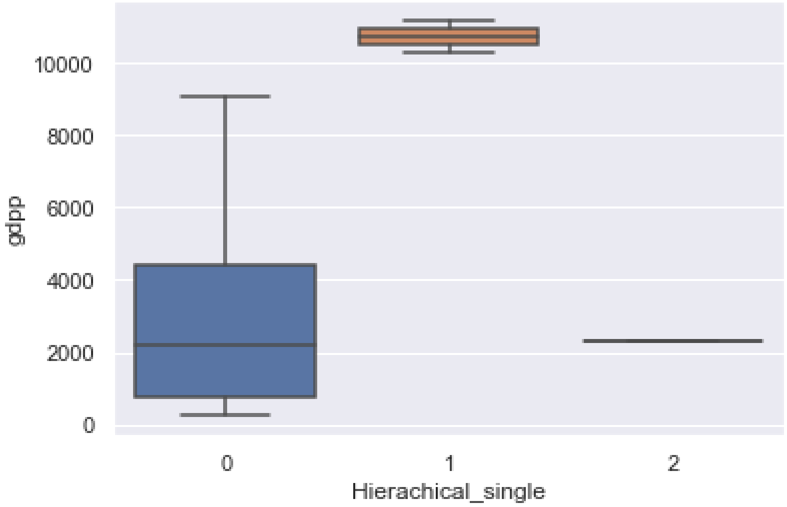


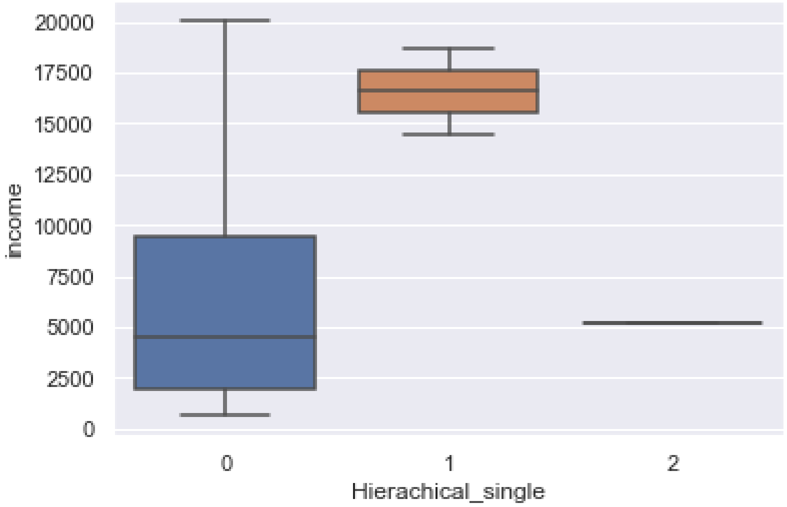
Single linkage Dendrogram

We can see that single linkage clustering is not promising since most of the data points are clustered in single cluster label

Clusters formed by Single linkage:

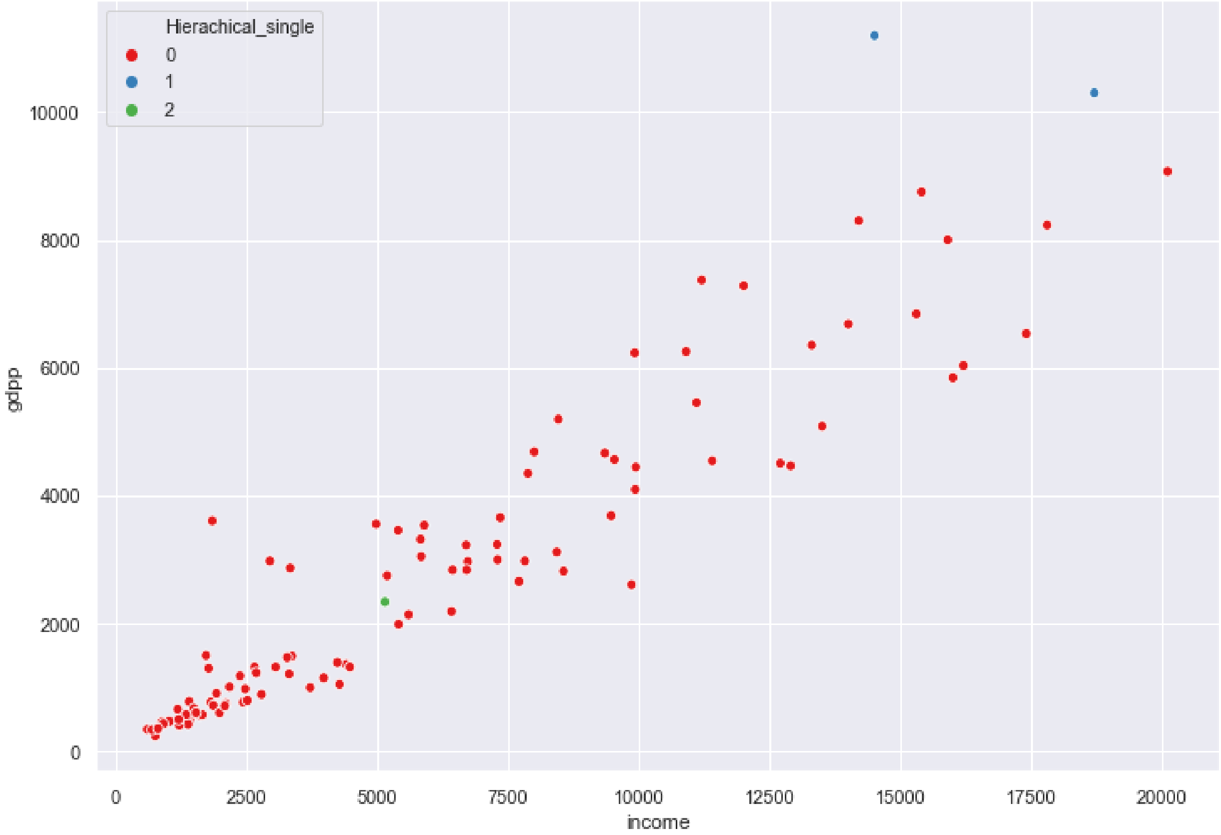


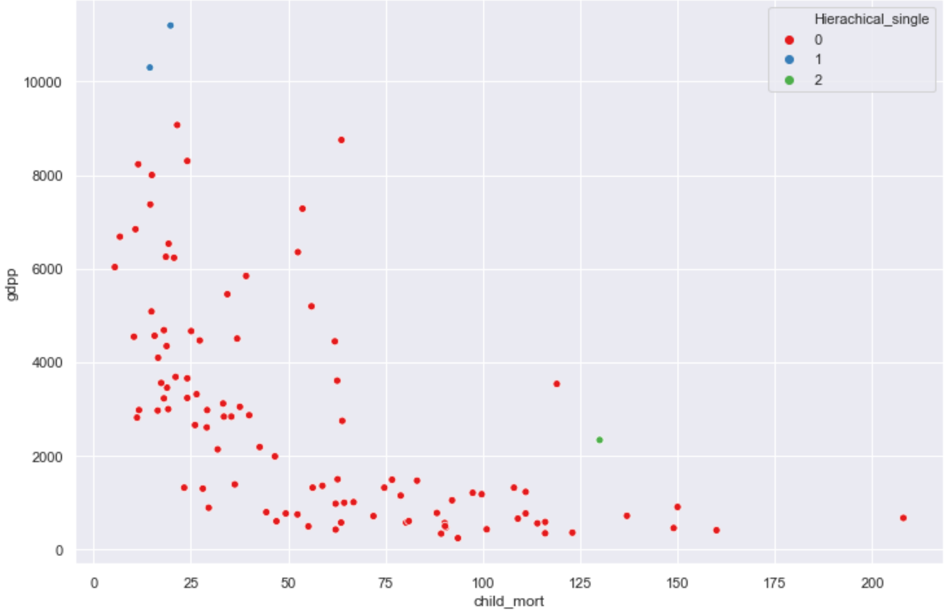


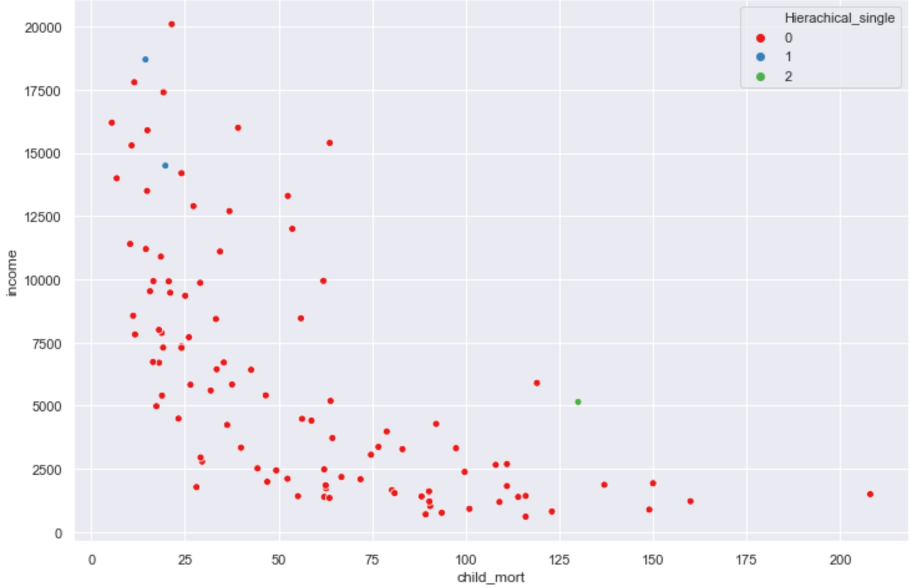


Clustering Model

Visualizing scatter plots for Hierarchical Single linkage:



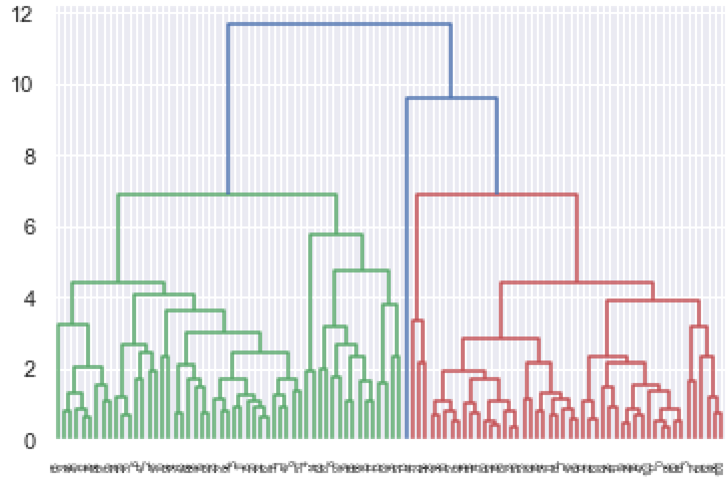




Clustering Model

Hierarchical clustering:

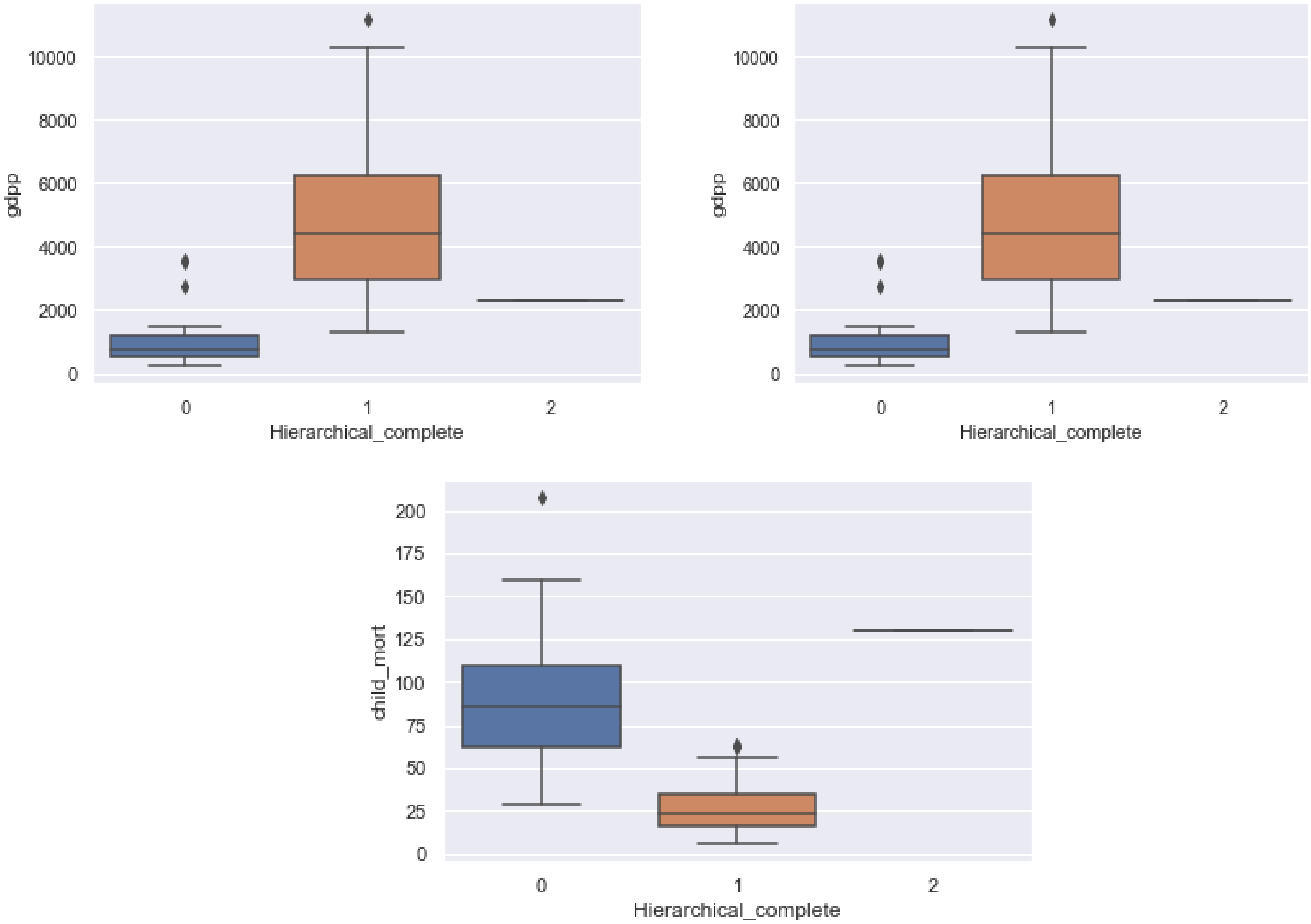
Complete linkage:



Complete linkage Dendrogram

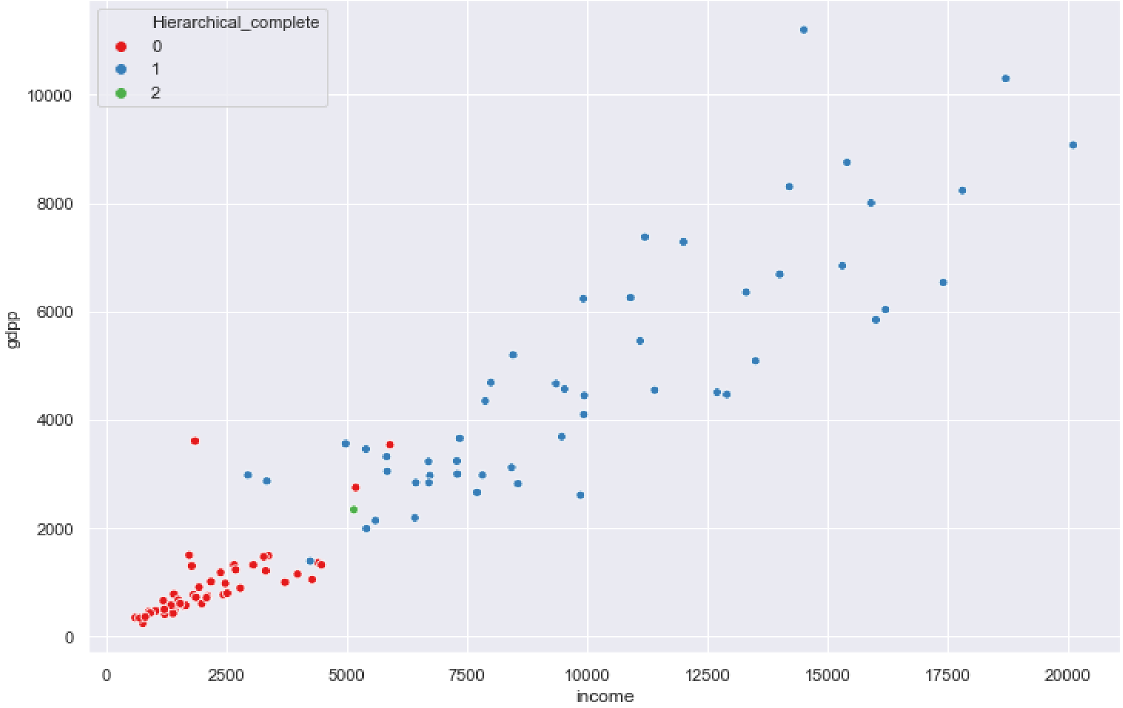
We can see that Complete linkage clustering is promising and we are able to find the backward countries from this cluster formation.

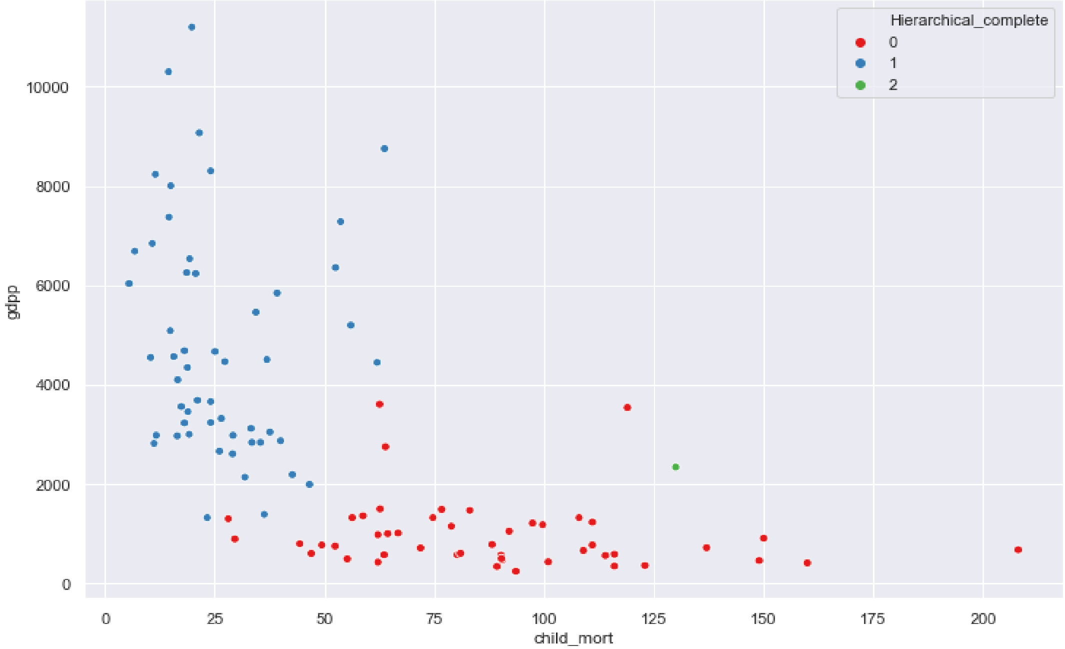
Clusters formed by Complete linkage:

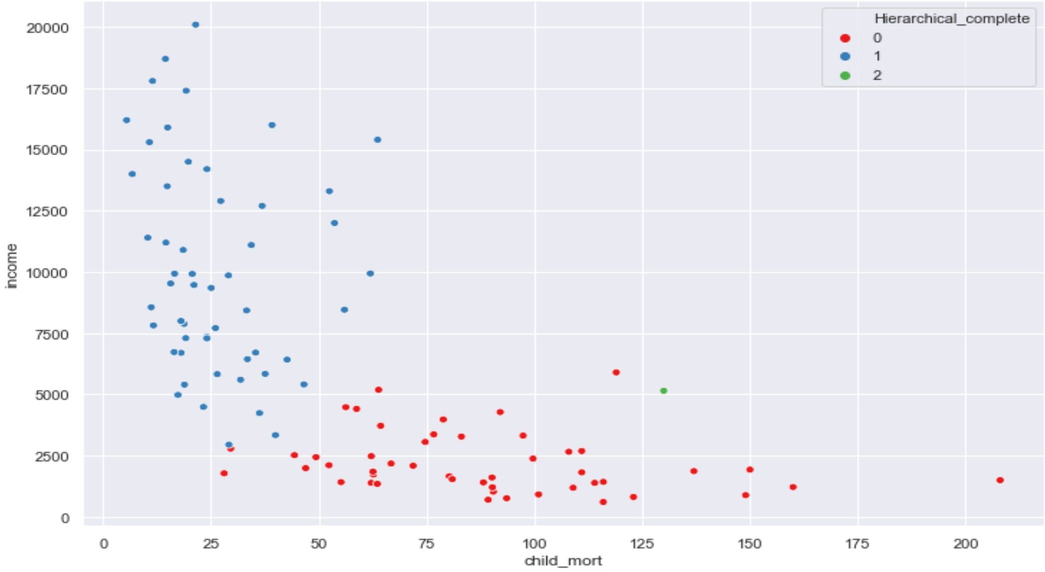


Clustering Model

Visualizing scatter plots for Hierarchical Complete linkage:







label = 0, low income, low GDP and high child mortality

label = 1, high income, high GDP and low child mortality

label = 2, moderate income, moderate GDP and high child mortality

Inferences and Recommendations

- In **k-means Clustering**, In order to suggest the backward countries, we should focus on data where **cluster label = 1**
- In **Hierarchical Clustering complete linkage**, In order to suggest the backward countries, we should focus on data where **cluster label = 0**
- After Checking common countries recommended by k-means and hierarchical clustering with low gdp, low income and high child mortality, we have almost **46 countries**, however, we need to suggest at least 5 countries from these clusters.
- *After sorting backward countries with lowest gdp, lowest income, high child_mort, final Top 10 backward countries are:*
- ***Burundi***
- ***Liberia***
- ***Congo, Dem. Rep.***
- ***Niger***
- ***Sierra Leone***
- ***Madagascar***
- ***Mozambique***
- ***Central African Republic***
- ***Malawi***
- ***Eritrea***

Thank You