Clustering Assignment

Shreyas R

Introduction

- ➢ **Problem Statement:**
- HELP International NGO has raised around $ 10 million and wants to use this money strategically and efficiently to fight poverty. For this they must know which countries are in the direst need of help.
- Therefore the countries are to be categorised based on the economic, social and health factors that determine the state of the country in terms of poverty.
- Finally a list of countries who are in the direst need of help have to be suggested to the CEO.
- ➢ **Available data:**
- “Country-data”, which contains the information about the economic, social and health conditions of different countries.

Approach and methodology

- **Data Inspection:**
  - Data cleaning
  - Univariate/Bivariate analysis
  - Outlier analysis
- **Clustering:**
  - Hopkins test
  - Scaling the data
  - To find optimal value of k: Silhouette analysis and elbow curve
  - K-means clustering
  - Cluster profiling: gdpp, child_mort and income
  - Hierarchical clustering(Single and complete linkage)
  - Final analysis and results

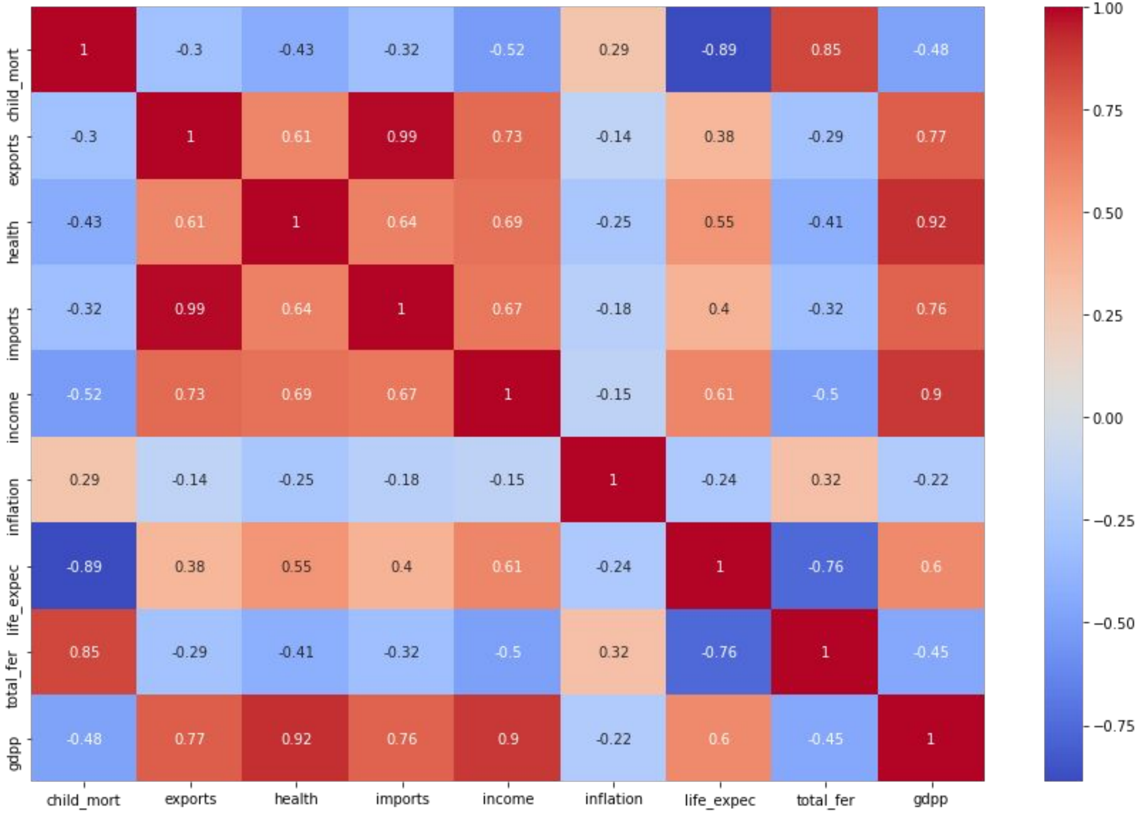
Correlations between variables

Income and health are highly correlated with GDP, import and exports are also highly correlated with GDP which is expected.

Imports is very highly correlated with exports.

Child Mortality is highly correlated with total fertility.

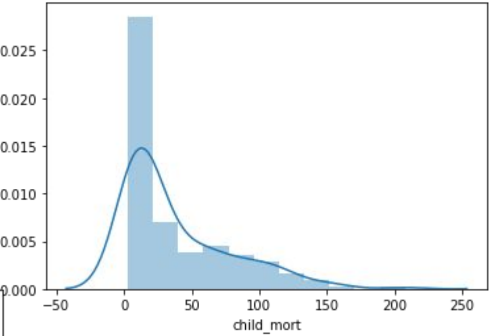
Life expectancy is highly negatively correlated with total fertility.

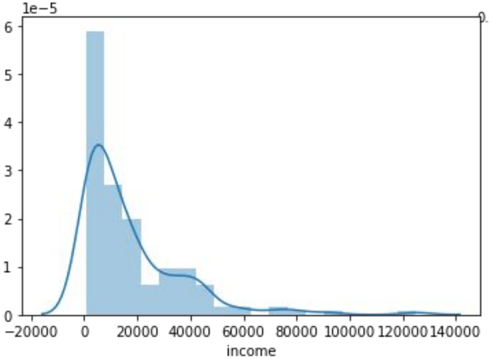


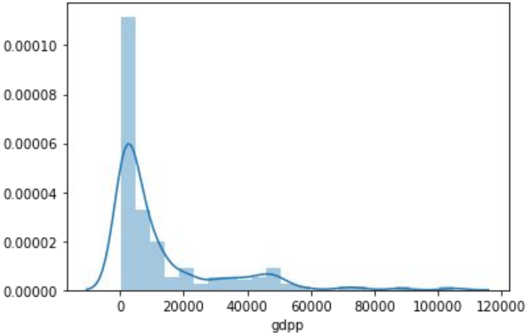
	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
child_mort	1	-0.3	-0.43	-0.32	-0.52	0.29	-0.89	0.85	-0.48
exports	-0.3	1	0.61	0.99	0.73	-0.14	0.38	-0.29	0.77
health	-0.43	0.61	1	0.64	0.69	-0.25	0.55	-0.41	0.92
imports	-0.32	0.99	0.64	1	0.67	-0.18	0.4	-0.32	0.76
income	-0.52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

Univariate analysis

Child Mortality, Income and GDP have a bimodal distribution, therefore it is highly likely that internal groups can be formulated.

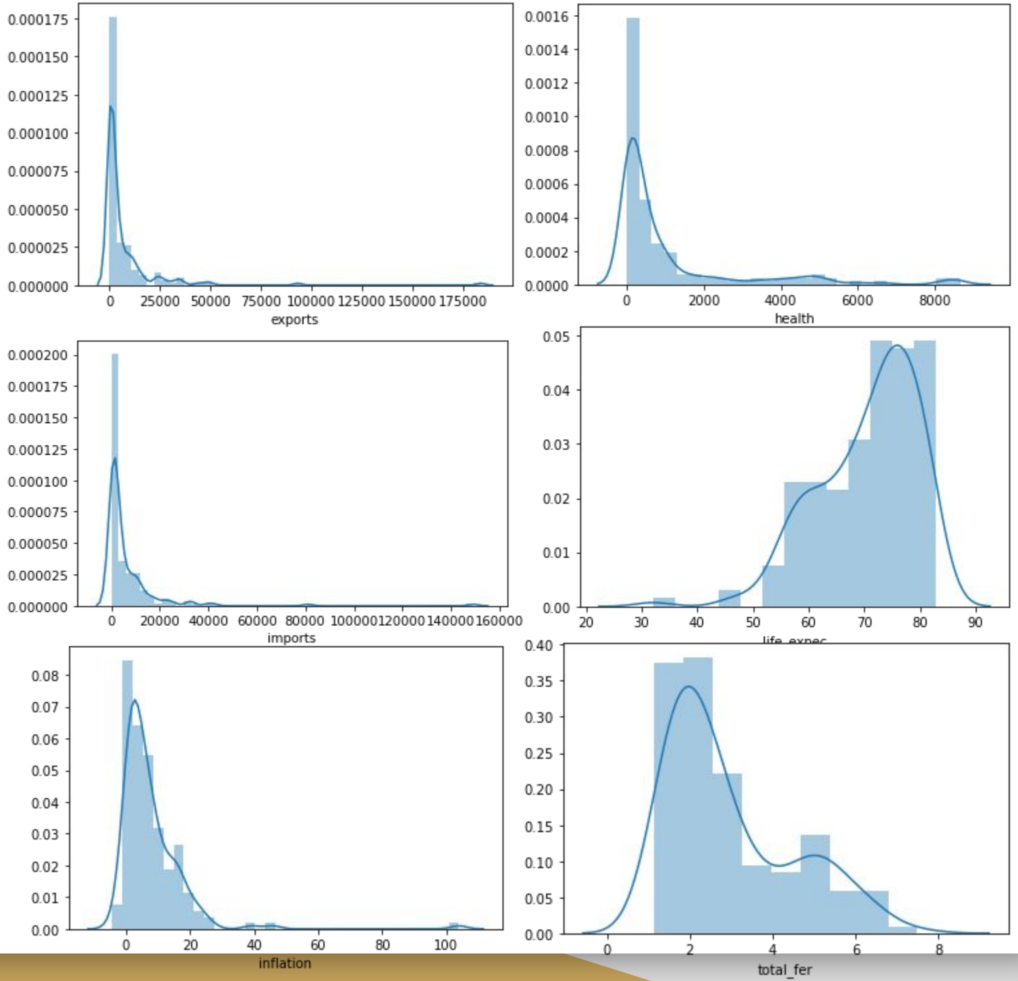




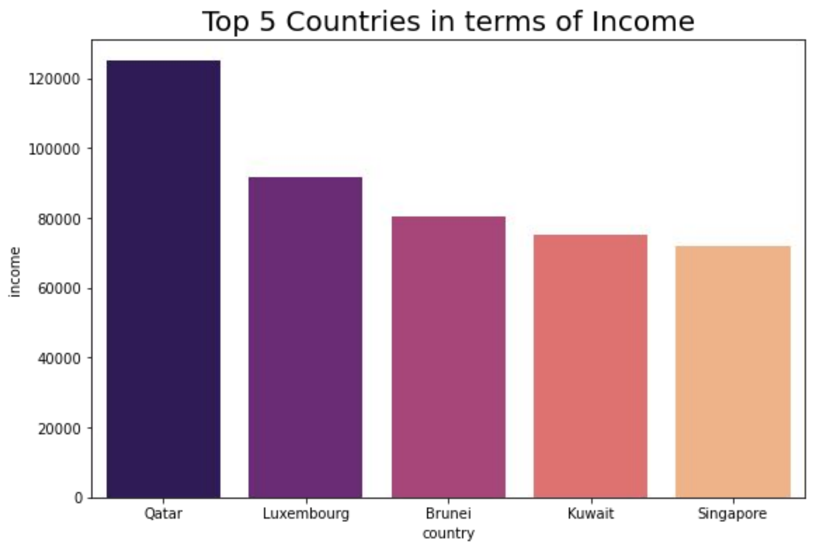


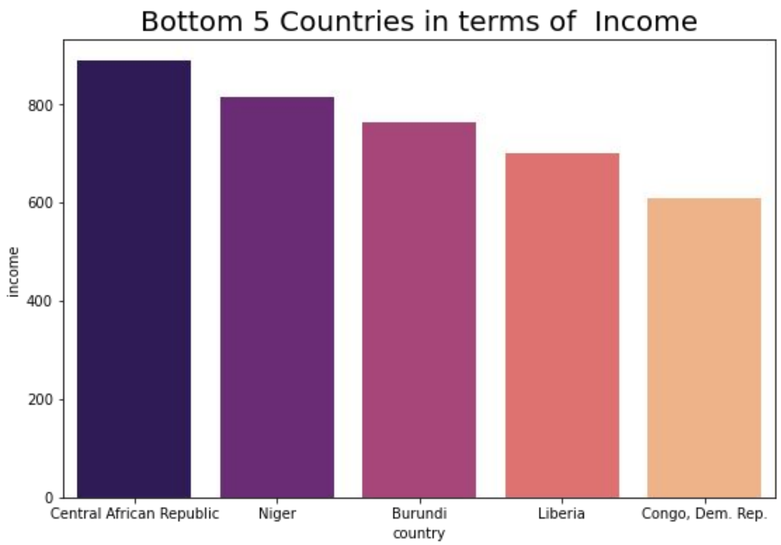
Univariate analysis

Exports, healths, imports, life expectancy, inflation and total fertility are normally distributed and are less likely to form internal groups.



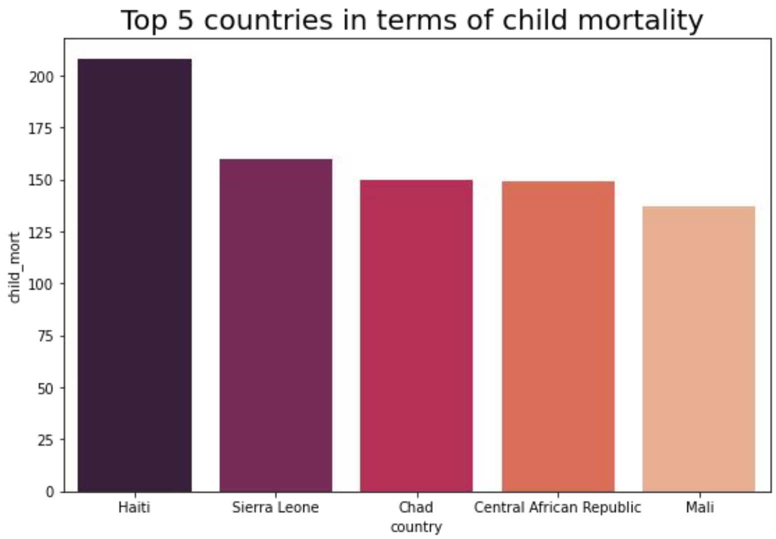
Bivariate analysis

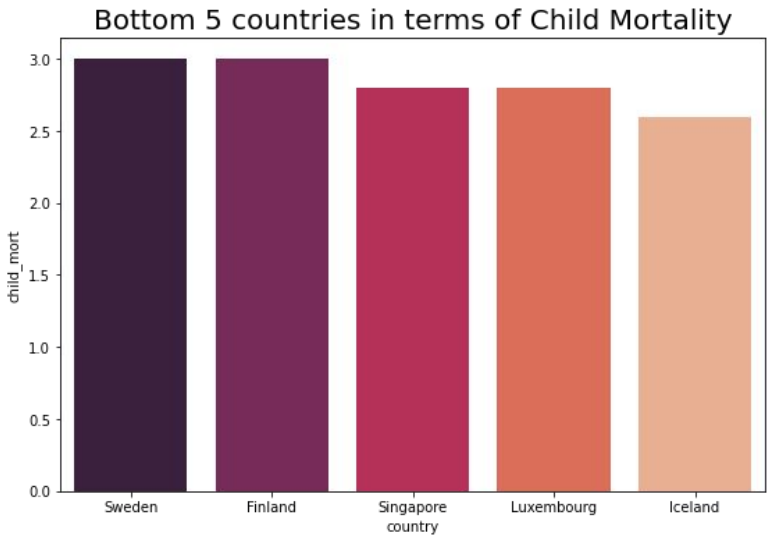




1) Qatar has an exceptionally high income, followed by Luxembourg.
2) Central African Republic, Niger , Burundi and Liberia belong to Africa, therefore most of the African countries have low income.

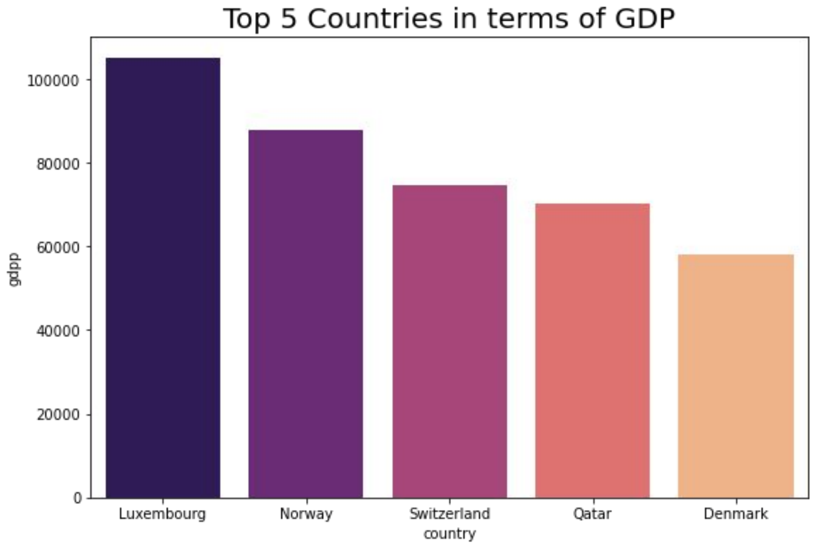
Bivariate analysis

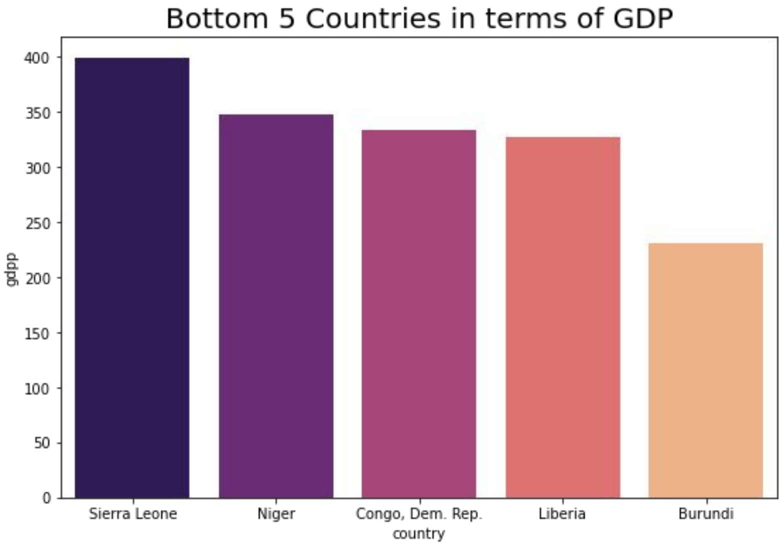

Top 5 countries in terms of child mortality


Bottom 5 countries in terms of Child Mortality

1) Haiti has a very high child mortality of around 200 for every 1000 children born, which means 20% of the children die within the age of 5, followed by Sierra Leone with 150.
2) Sweden, Finland, Singapore and Luxembourg have a very low child mortality of below 3.

Bivariate analysis



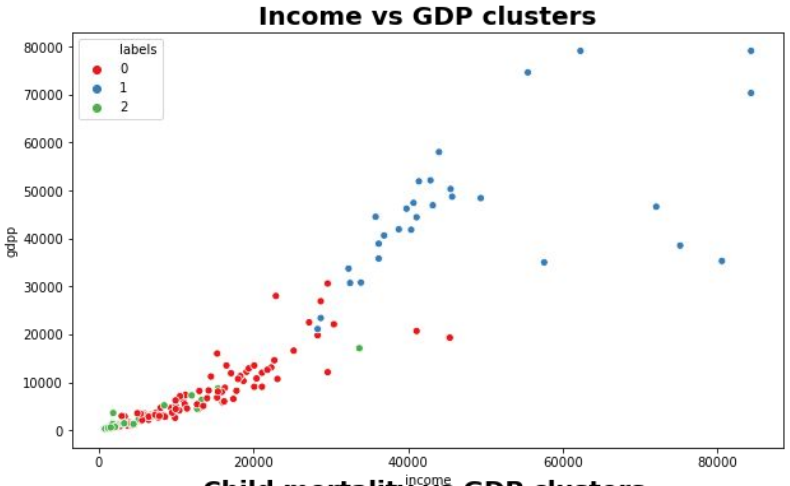


1) We see almost similar stats for GDP compared to income, Luxembourg having the highest GDP.
2) As expected we see the African countries with lowest GDP's.

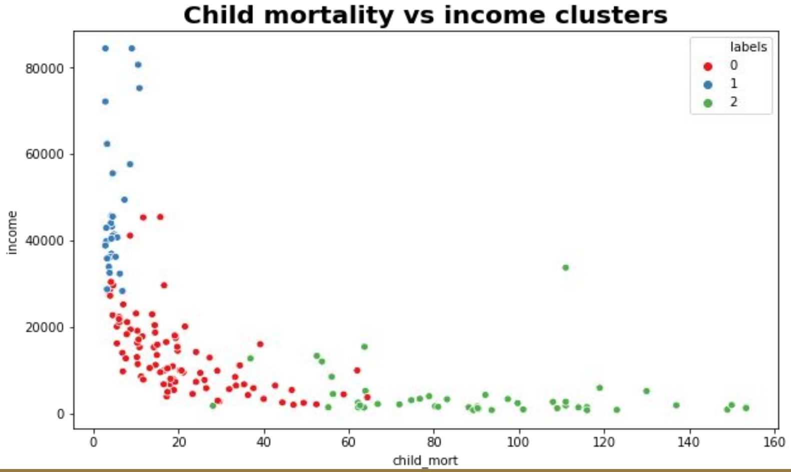
Clusters formed from k-means clustering

The three clusters can be distinguished well in all the scatter plots.

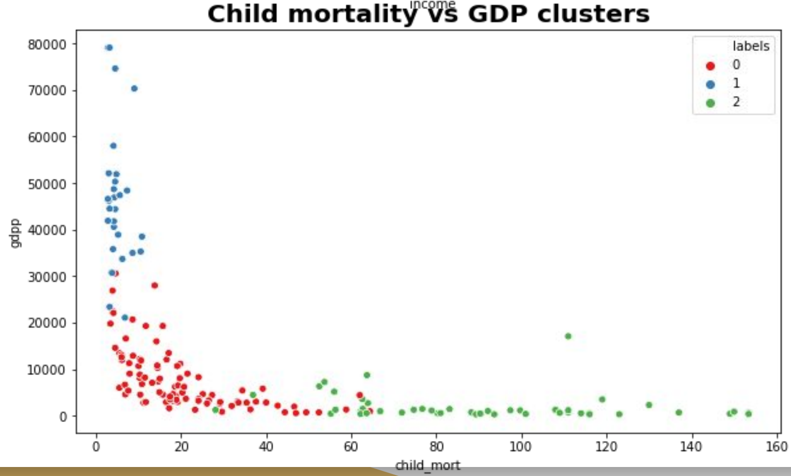
We can see that cluster 2 (green points) has high child mortality, low income and low GDP, therefore the countries in this cluster are in dire need of help.



Income vs GDP clusters

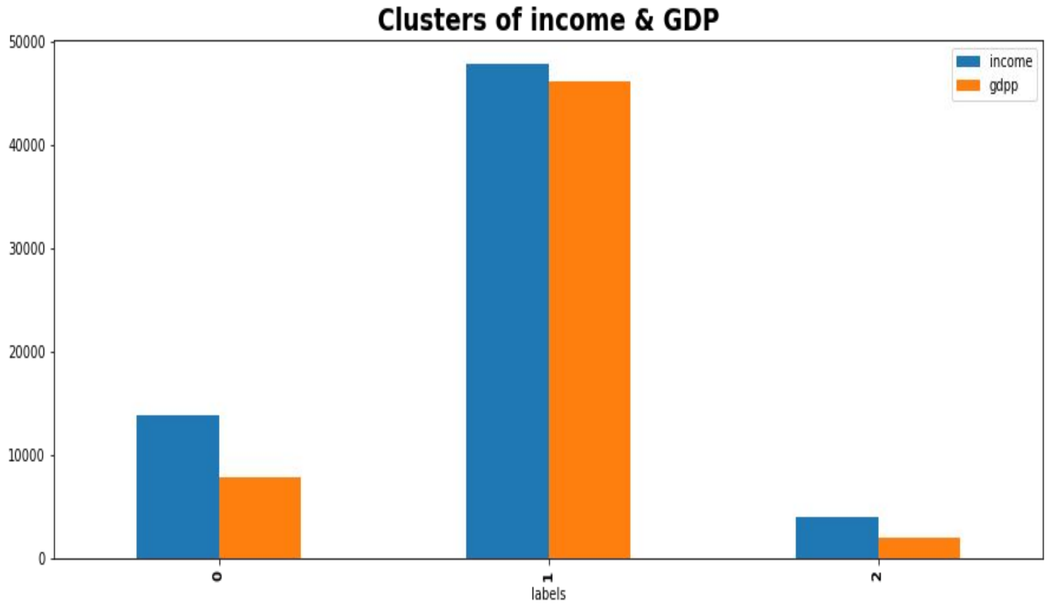


Child mortality vs income clusters



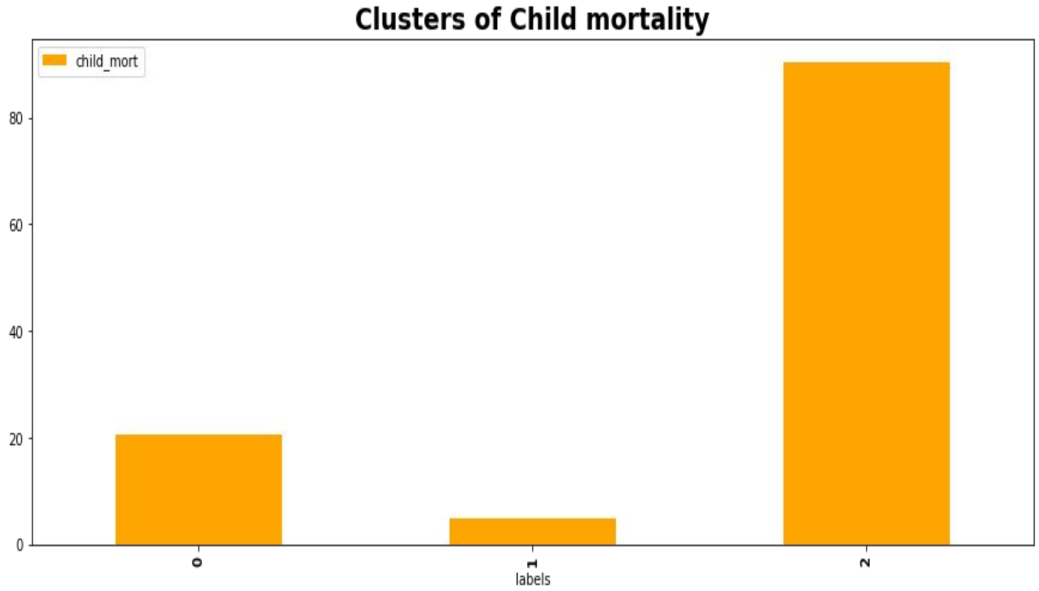
Child mortality vs GDP clusters

K means clustering results(Income and GDP)



We see that cluster 2 has low income and GDP , therefore the countries which belong to this cluster are the ones to focus on.

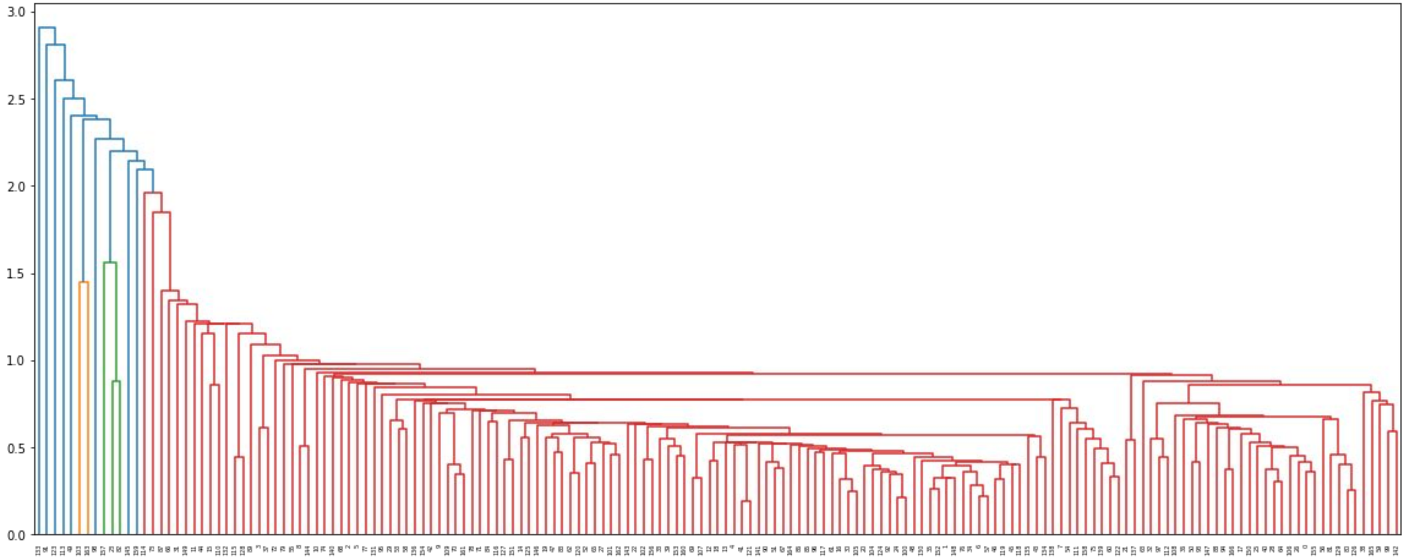
K means clustering results(Child Mortality)



We see that cluster 2 has high Child Mortality, therefore the countries which belong to this cluster are the ones to focus on.

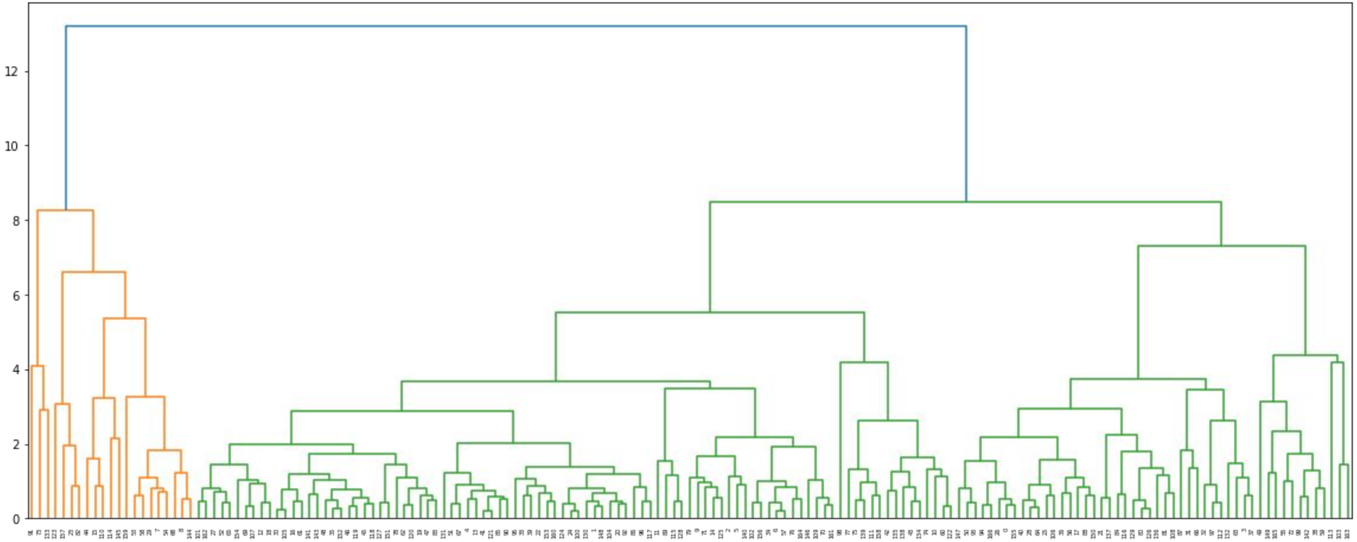
Countries which have very low income and GDP and also have high Child Mortality are **Sierra Leone, Haiti, Chad, Central African Republic, Mali, Nigeria, Niger, Angola, Congo. Dem. Rep. and Burkina Faso**

Hierarchical clustering(Single linkage)



It is clearly evident that single linkage does not provide good results, therefore we need to proceed with complete linkage.

Hierarchical clustering(Complete linkage)

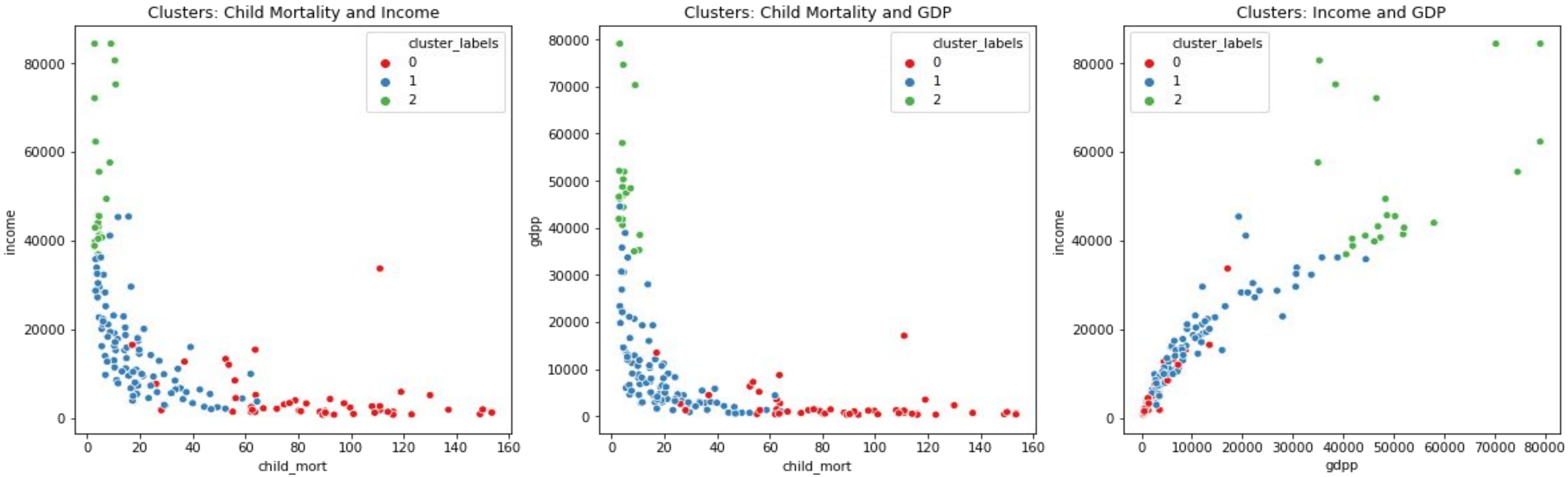


In complete linkage dendrogram it looks like the clusters are evident and can be analysed easily.

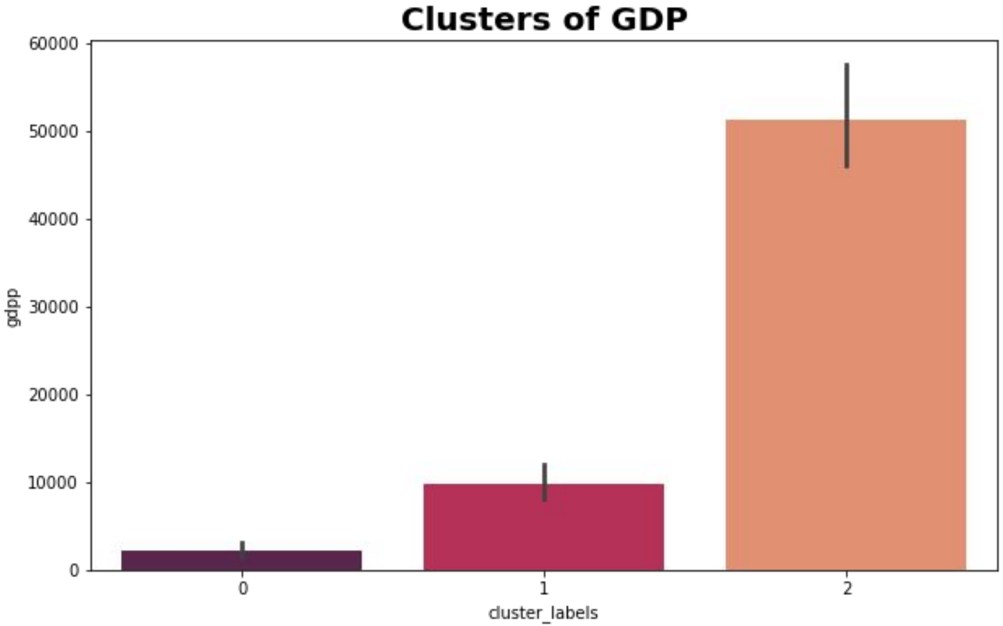
Clusters formed from Hierarchical clustering

We can see that the 3 clusters can be distinguished pretty well in all the scatter plots.

We can see a similar distribution for both k means and hierarchical clustering.

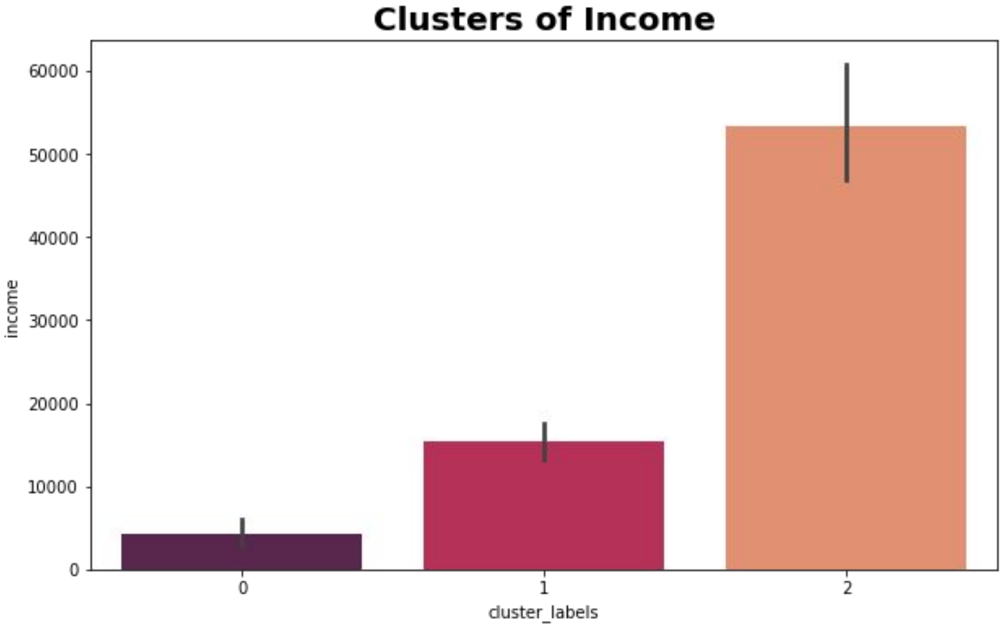


Hierarchical clustering results(GDP)



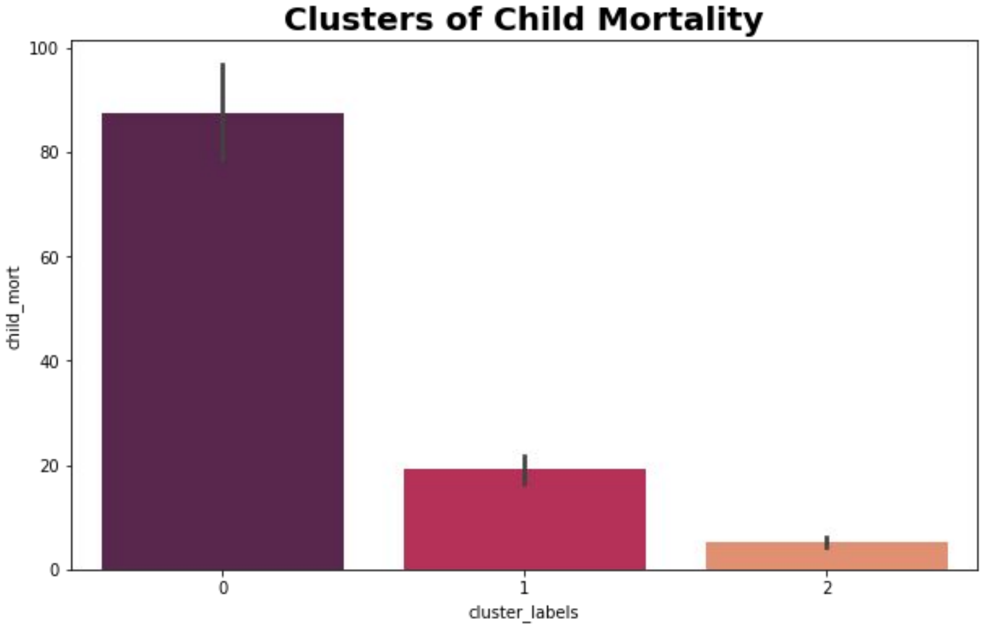
We see that cluster 2 has low GDP , therefore the countries which belong to this cluster are the ones to focus on.

Hierarchical clustering results(Income)



We see that cluster 2 has low GDP , therefore the countries which belong to this cluster are the ones to focus on.

Hierarchical clustering results(Child Mortality)



We see that the same cluster 0 which had low income and GDP also has high child mortality, similar to k-means clustering results. Therefore the countries belonging to this cluster are the ones in dire need of aid.

And these countries are **Sierra Leone, Haiti, Chad, Central African Republic, Mali, Nigeria, Niger, Angola, Congo. Dem. Rep. and Burkina Faso**

Conclusion

- Income, GDP and Child mortality are the most important factors.
- Countries which have low income also have low GDP, which is resulting in high child mortality.
- We got similar looking clusters for both k means and hierarchical clustering, the number of values in each cluster is more even in k-means clustering, therefore it is more reliable.
- The child mortality increases with increase in total fertility
- The top 10 countries in the direst need to help are : **Sierra Leone, Haiti, Chad, Central African Republic, Mali, Nigeria, Niger, Angola, Congo. Dem. Rep. and Burkina Faso**.
- Most of these countries belong to Africa, therefore Africa in general has to be focused more.