Countries in need of immediate aid - Clustering assignment

Shyamala Rajasekar

Problem Statement

An NGO called HELP intends to help countries with basic amenities and fighting poverty. It has raised around 10 million dollars. The NGO wants to identify 5 countries that are in need of immediate aid based on parameters such as Income, Child Mortality and GDPP.

Objective

- The objective of this analysis is to cluster the countries based on parameters such as **income, GDPP and Child Morality** in order to find a cluster of countries struggling in terms of these basic parameters.
- **Further from the cluster of countries with low GDPP, low Income and high Child mortality, this analysis should be able to present 5 countries in need of immediate help.**

Analysis

We are given a dataset with 166 countries and 9 parameters. We have further divided these into two broad categories

Positive parameters : Increase in these indicate development/progress

1. Income
2. Exports
3. Imports
4. Health
5. GDPP

Negative parameters : Increase in these parameters indicate poverty or lack of basic amenities

1. Child mortality
2. Inflation
3. Life expectancy
4. Total fertility

1. Data cleaning

- There were no null values in the data set
- Data types of the parameters were correct
- There were no duplicate records of the country.
- Health, Imports and exports are as a percentage of GDPP. They were converted to their actual figures.

  Ex. Health expenditure in Afghanistan was 7.58% of GDPP

Outlier treatment

Capping the upper outliers in the below parameters

- Exports, Health, income, GDPP, life expectancy, Imports are the progress indicators of a country.
- The countries which have higher percentage of the above parameters are developed and self sufficient and not in need of aid.
- Hence those countries with more than 99th percentile of these parameters were capped at 99th percentile.

Capping the lower outliers in the below parameters:

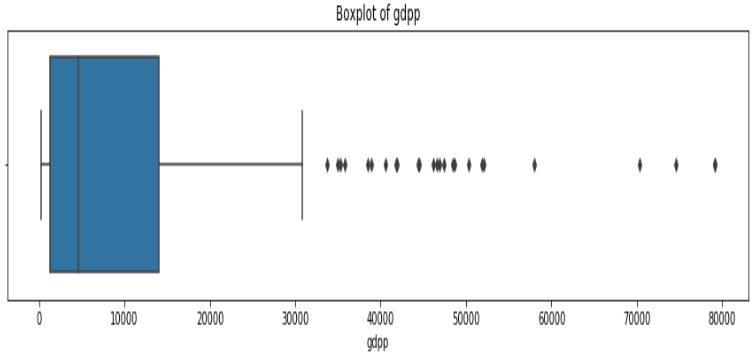
- **inflation** : The measurement of the annual growth rate of the Total GDP. Governments often strive for 2-3% inflation rates. High inflation can cause the economy to decline. Those with higher inflation needs aid.
- **Total fertility** : This must be around 2.1.With rise in TFR, Human development index decreases. Higher the TFR, higher need of aid.
- **Child Mortality** : It is the death of children under 5 years of age per 1000 live births. Higher child mortality indicates that the country might need aid to uplift its healthcare.

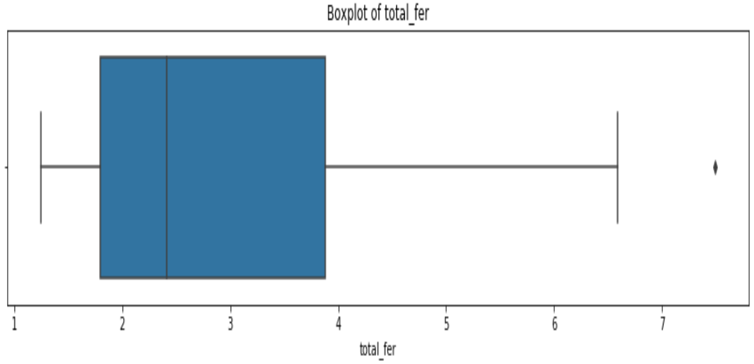
Parameters	Before outlier treatment range	After Outlier treatment range
Child Mortality	2.6 - 208	2.8 to 208
Income	609 to 125000	609 to 84374
Health	12 to 8663	78 to 8410
Exports	1.07 to 183750	1.07 to 64794.26
Imports	0.6 to 149100	55371 to 149100
GDPP	231 to 105000.	231 to 79088
Life expectancy	32 years to 82 years	32 years to 80 years
Total fertility	1.15 to 7.49	1.24 to 7.49
Inflation	-4.2 to 104	-2.34 to 104

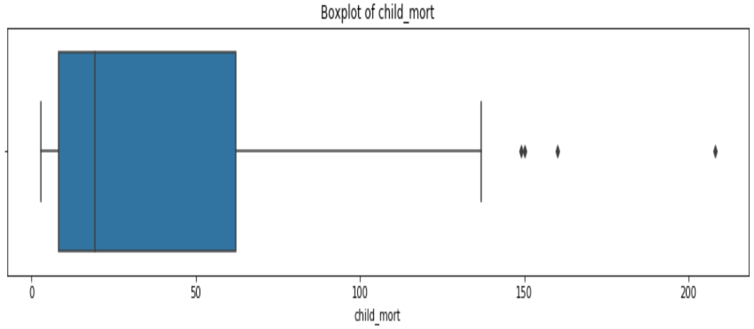
Univariate analysis

Conclusions drawn from univariate analysis :

1. There are a lot of countries beyond the 75th percentile of exports. Median is around 2000.
2. Child mortality ranges from 2 to 208 with median at 20.
3. Fertility rates range from 1 to 6.8 with median at 2.5
4. GDPP 75th percentile is at 30000 and ranges upto 80000.



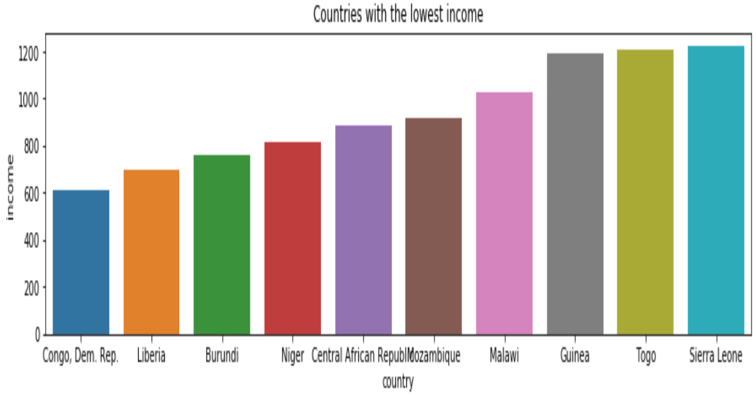


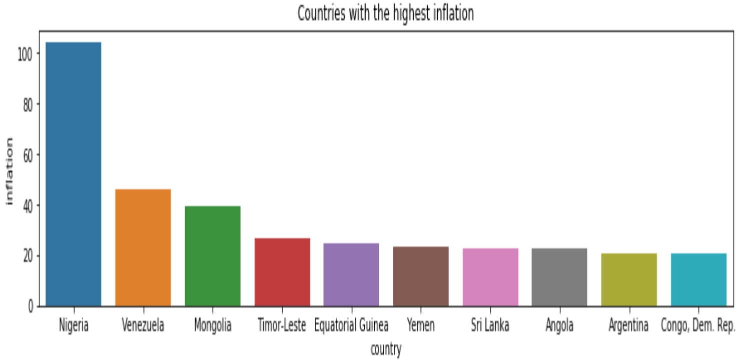


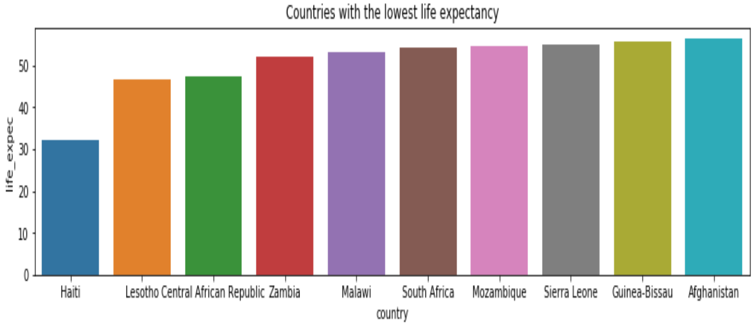
BIVARIATE ANALYSIS

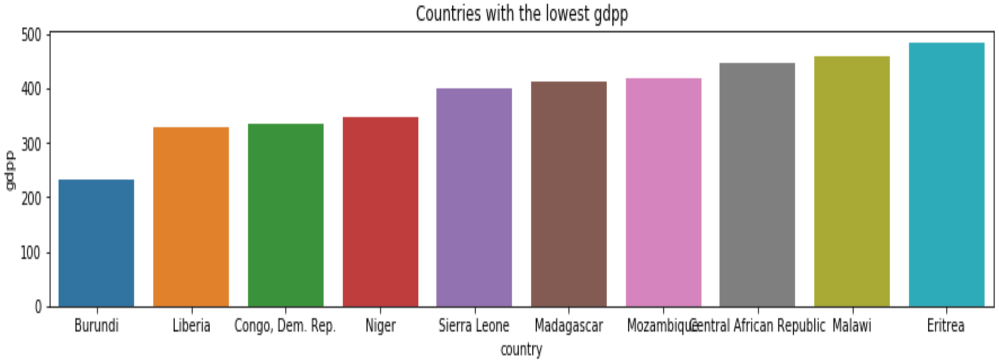
From bivariate analysis, the following conclusions were drawn:

1. Haiti has the highest child mortality rate
2. Congo Dem republic has the lowest income
3. Nigeria has the highest inflation
4. Haiti has the lowest life expectancy
5. Afganistan has the lowest amount of imports.
6. Eritrea spends the lowest in health.
7. Niger has the highest fertility rate.
8. Burundi has the lowest GDPP
9. Myanmar exports the lowest.









Correlation

- Child Mortality and income are negatively correlated.
- Child Mortality and life expectancy are negatively correlated.
- Child Mortality and GDPP are negatively correlated.
- Income and total fertility are negatively correlated.
- life expectancy and total fertility are negatively correlated.

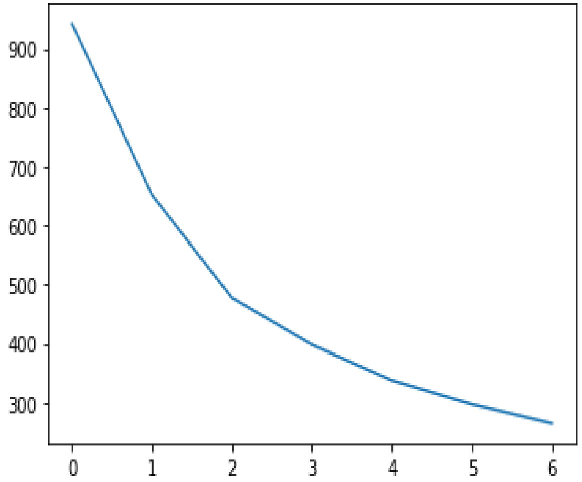
2. Data Preparation :

- Hopkins score is 0.85 for several iterations. Hence the dataset is fit for clustering
- All variables are numerical variables. They were all scaled.

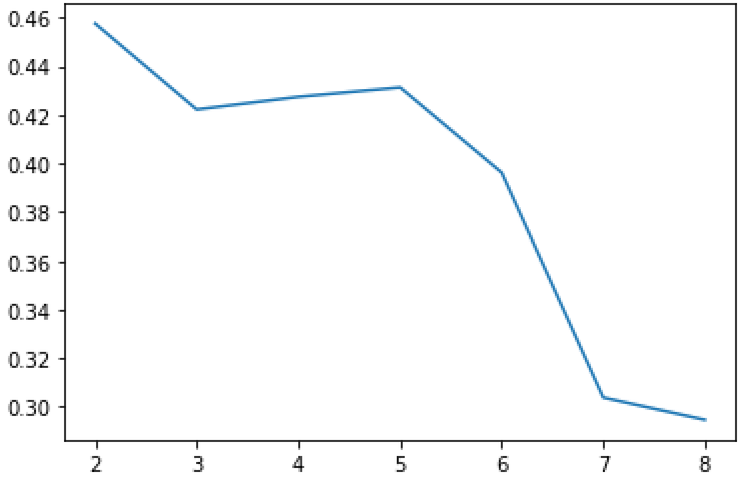
3. Clustering

- K-means clustering method was used to cluster the countries.
- For k-means, the value of k is pre defined. It was determined through

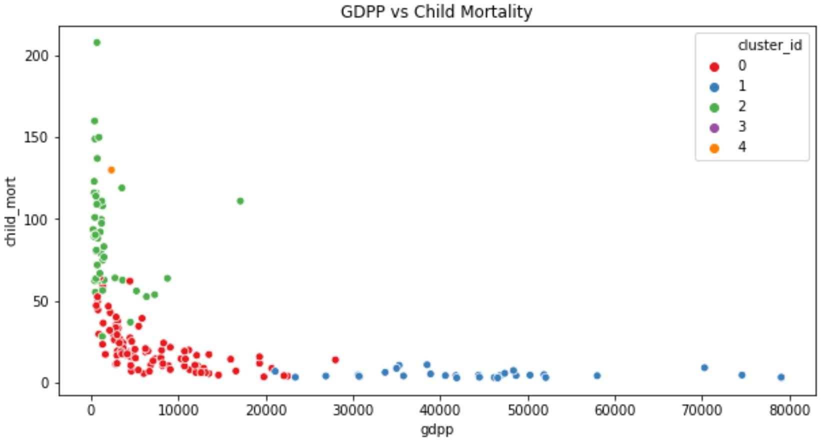
1. Elbow curve

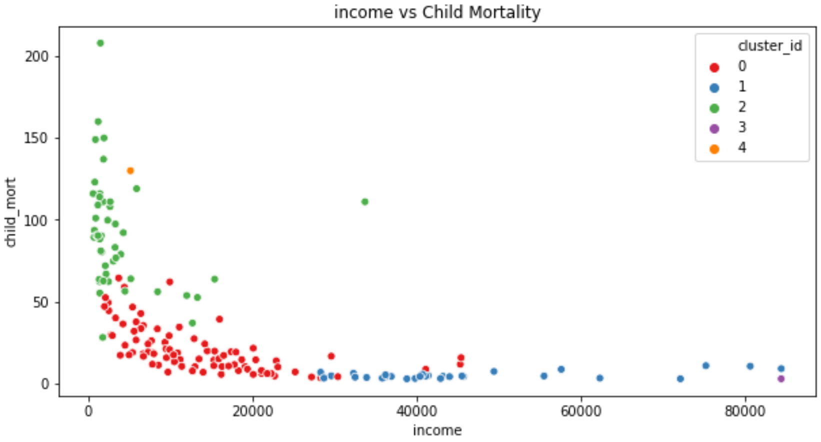


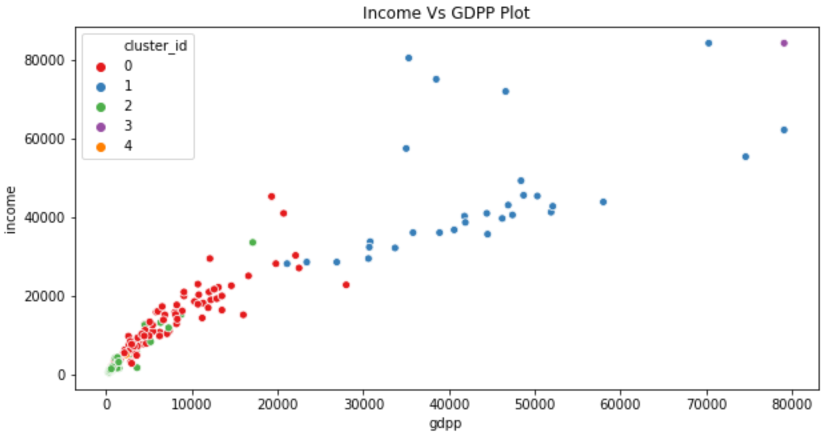
2. Silhouette score



K-means checked for 5 and 6 clusters. 6 clusters did not yield satisfactory results. Hence k value was taken as 5 for further analysis.



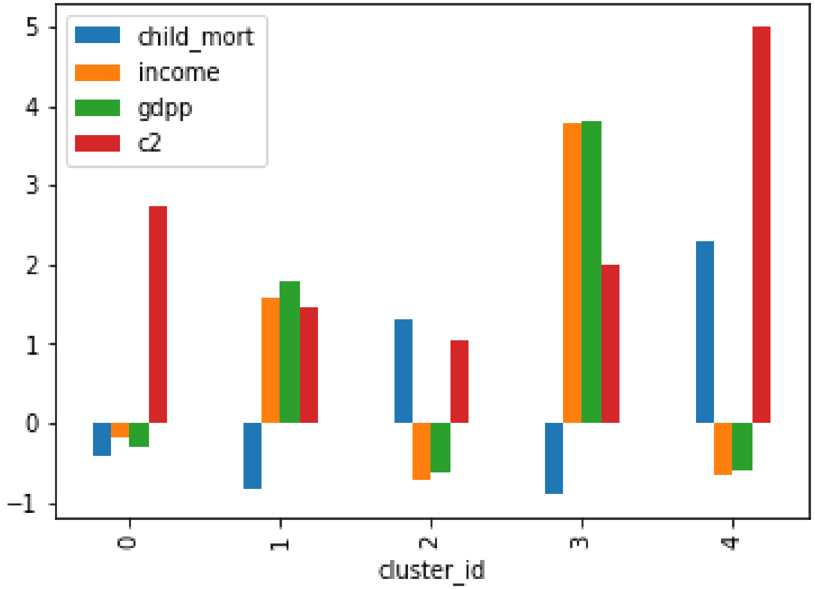




Graphs were plotted for clusters based on child mortality, income and GDPP and the following deductions were made :

- **Cluster 2** has the highest child mortality and the lowest income.
- **Cluster 1** has the lowest child mortality and highest income.
- **Cluster 2** has the child mortality and low GDPP.
- **Cluster 1** has the highest GDPP and lowest child mortality.
- **Cluster 3** has the highest GDPP and Income.
- **Cluster 2** has the lowest GDPP and income.

Results from K-means



The countries that need immediate aid

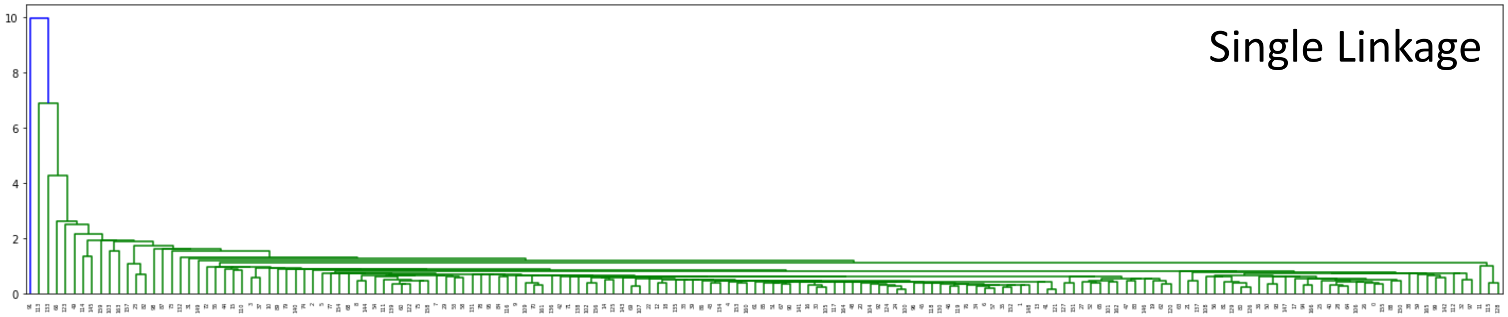
- Nigeria
- Haiti
- Sierra leone
- Chad
- Central African republic
- Mali

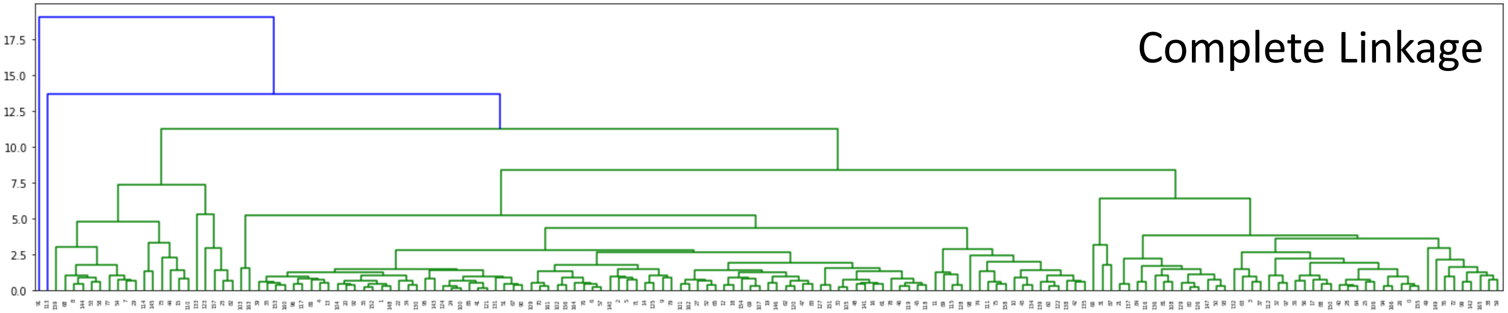
Cluster Profiles derived from K-means

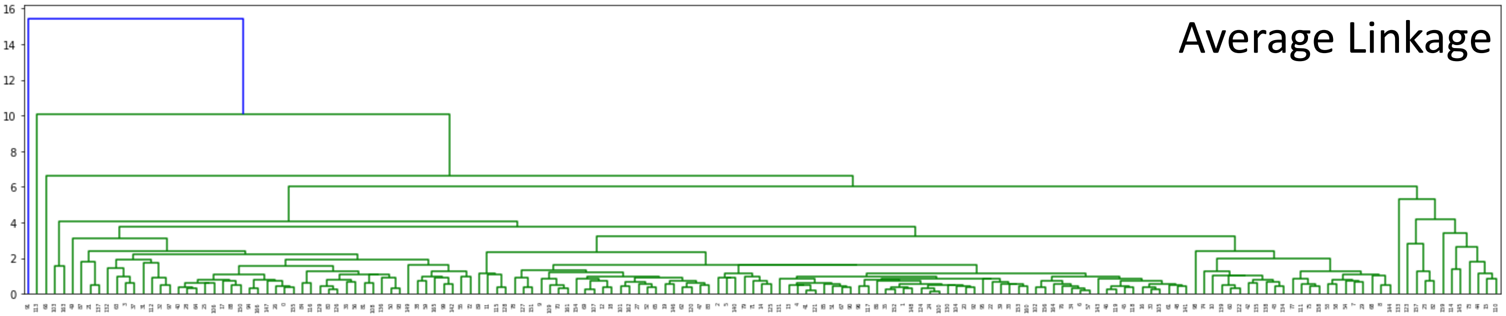
- **Cluster 0** has low child mortality, low-income, low gdpp
- **Cluster 1** has low child mortality, high income and high gdpp
- **Cluster 2** has low income, high child mortality, low gdpp.
- **Cluster 3** has high income, high gdpp, low child mortality.
- **Cluster 4** has high child mortality, low income, and low GDPP. This is the cluster that need immediate aid followed by cluster 2.

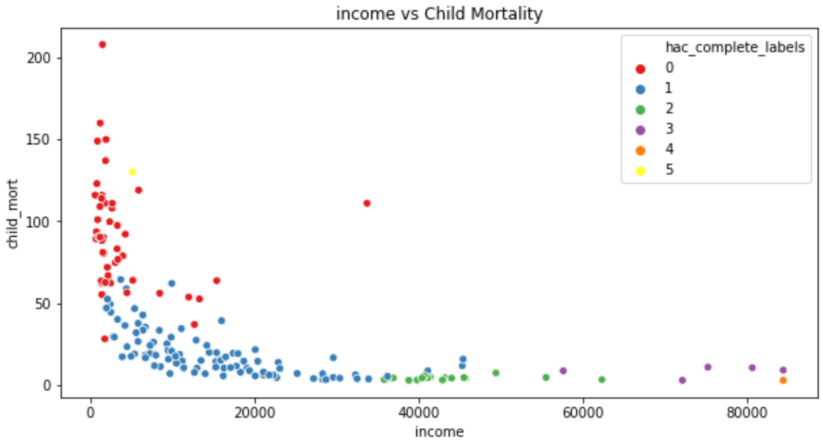
Hierarchical Clustering

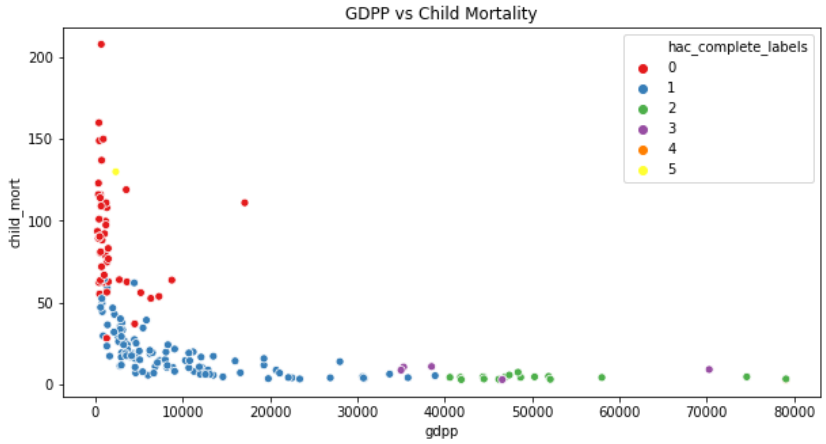
Three different linkages were used to determine the number of clusters.

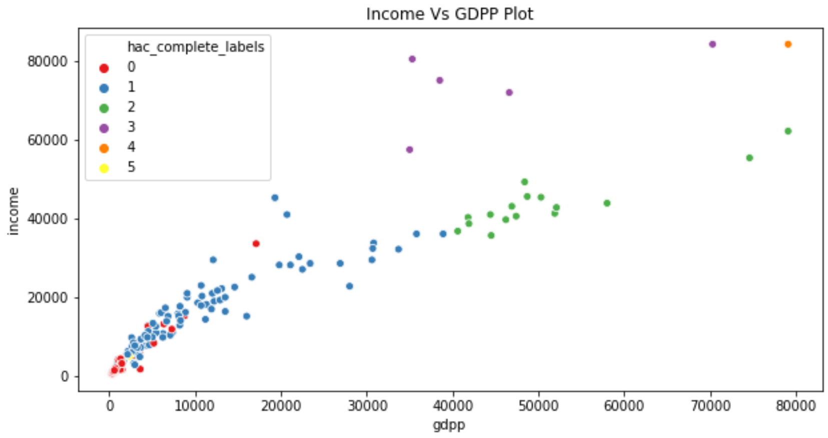







income vs Child Mortality


GDPP vs Child Mortality


Income Vs GDPP Plot

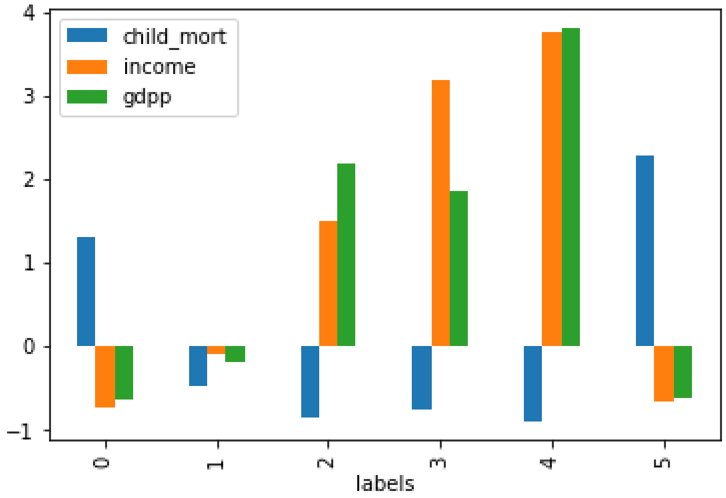
Upon cutting the dendrogram derived from complete linkage, 6 clusters were obtained. Labels were generated for 6 clusters and clustering was done.

Graphs were plotted for three main parameters : **Child mortality, Income, GDPP**

Deductions made:

- **Cluster 2** has high GDPP and high income.
- **Cluster 4** has the highest GDPP and high income.
- **Cluster 0** has the lowest income and GDPP.
- **Cluster 0** has the highest child mortality and lowest GDPP
- **Cluster 5** also has high child mortality.
- **Cluster 0** has high child mortality and low income.
- **Cluster 5** has high child mortality and low income.

Results from Hierarchical Clustering



Countries in need of immediate aid

1. Nigeria
2. Haiti
3. Sierra leone
4. Chad
5. Central African Republic.

Clustering is done with child mortality as the priority followed by GDPP and then income

Cluster profiles :

- **Cluster 0** : High child mortality, low income, low GDPP
- **Cluster 1** : Low child mortality, low income, low GDPP
- **Cluster 2** : Low child mortality, high income, high GDPP
- **Cluster 3** : Low child mortality, significantly high income and GDPP
- **Cluster 4** : Low Child mortality and very high income and GDPP. This cluster seems consists of developed countries
- **Cluster 5** : High Child mortality and low income and GDPP. This is the cluster in need of immediate aid.

COUNTRIES IN NEED OF IMMEDIATE AID :

1. Nigeria
2. Haiti
3. Sierra leone
4. Chad
5. Central African Republic.