CLUSTERING ASSIGNMENT

PREPARED BY

PRITAMKUMAR SURYAVANSHI

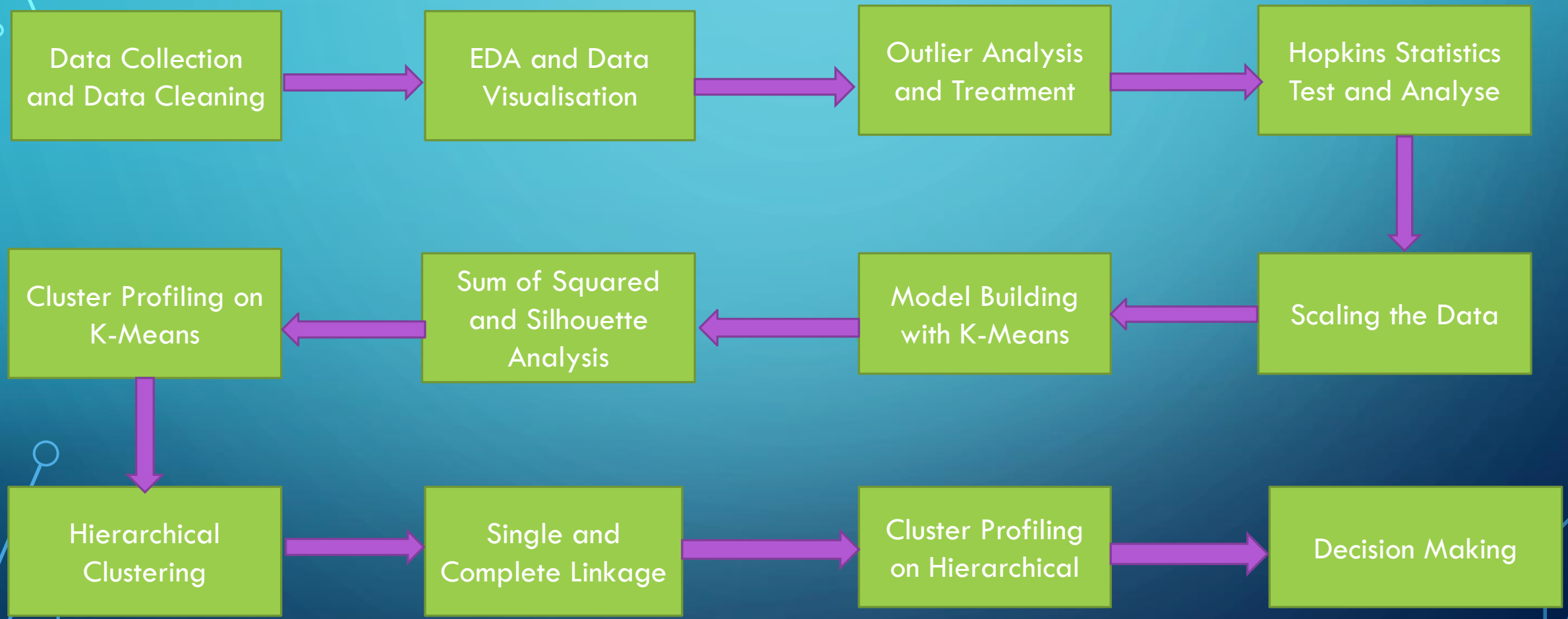
OBJECTIVE

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. IT RUNS A LOT OF OPERATIONAL PROJECTS FROM TIME TO TIME ALONG WITH ADVOCACY DRIVES TO RAISE AWARENESS AS WELL AS FOR FUNDING PURPOSES.
AFTER THE RECENT FUNDING PROGRAMS, THEY HAVE BEEN ABLE TO RAISE AROUND $10 MILLION. NOW THE CEO OF THE NGO NEEDS TO DECIDE HOW TO USE THIS MONEY STRATEGICALLY AND EFFECTIVELY. THE SIGNIFICANT ISSUES THAT COME WHILE MAKING THIS DECISION ARE MOSTLY RELATED TO CHOOSING THE COUNTRIES THAT ARE IN THE DIREST NEED OF AID.

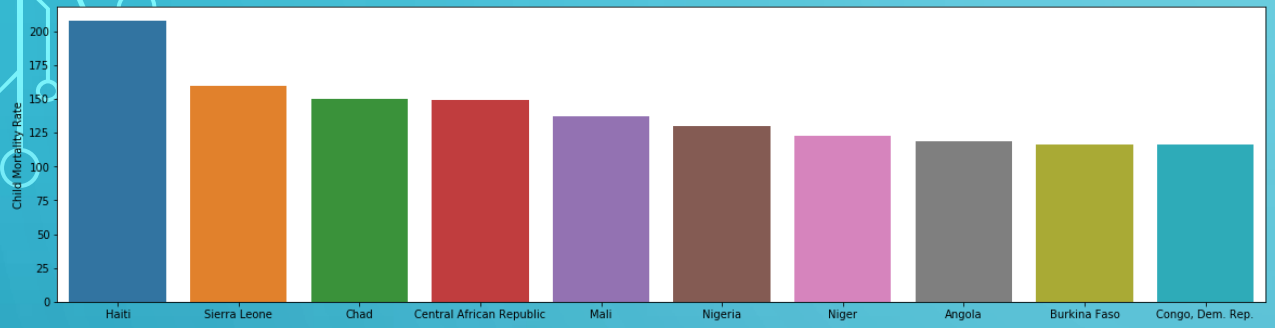
BUSINESS GOAL

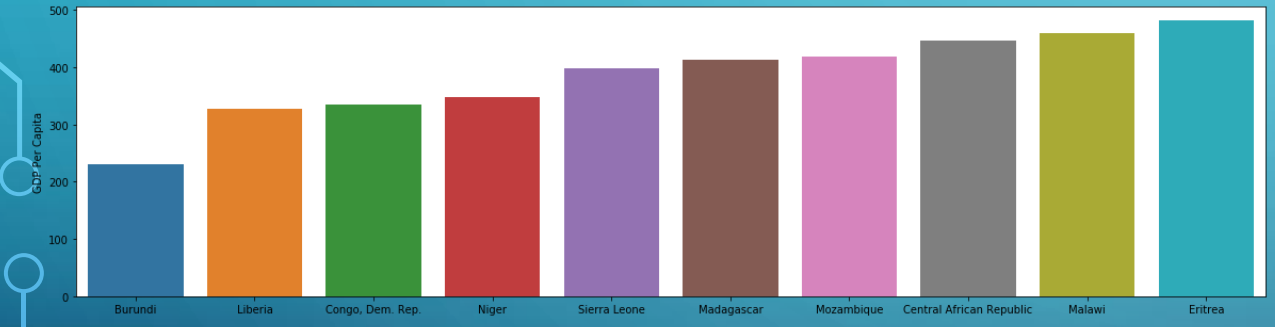
Identify top countries that are direst need of aid. Your job is to categories the countries using some socioeconomic and health factors that determine the overall development of the country. Then you need to suggest the countries which the CEO needs to focus on the most.

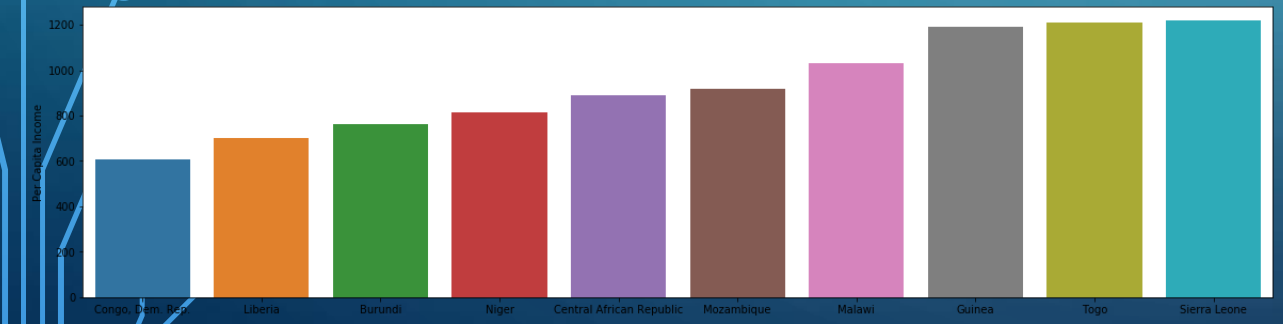
ANALYSIS METHODOLOGY



EDA AND DATA VISUALISATION







1) Child Mortality – List of countries that defines death of children under 5 years of age per 1000 live births in descending order.

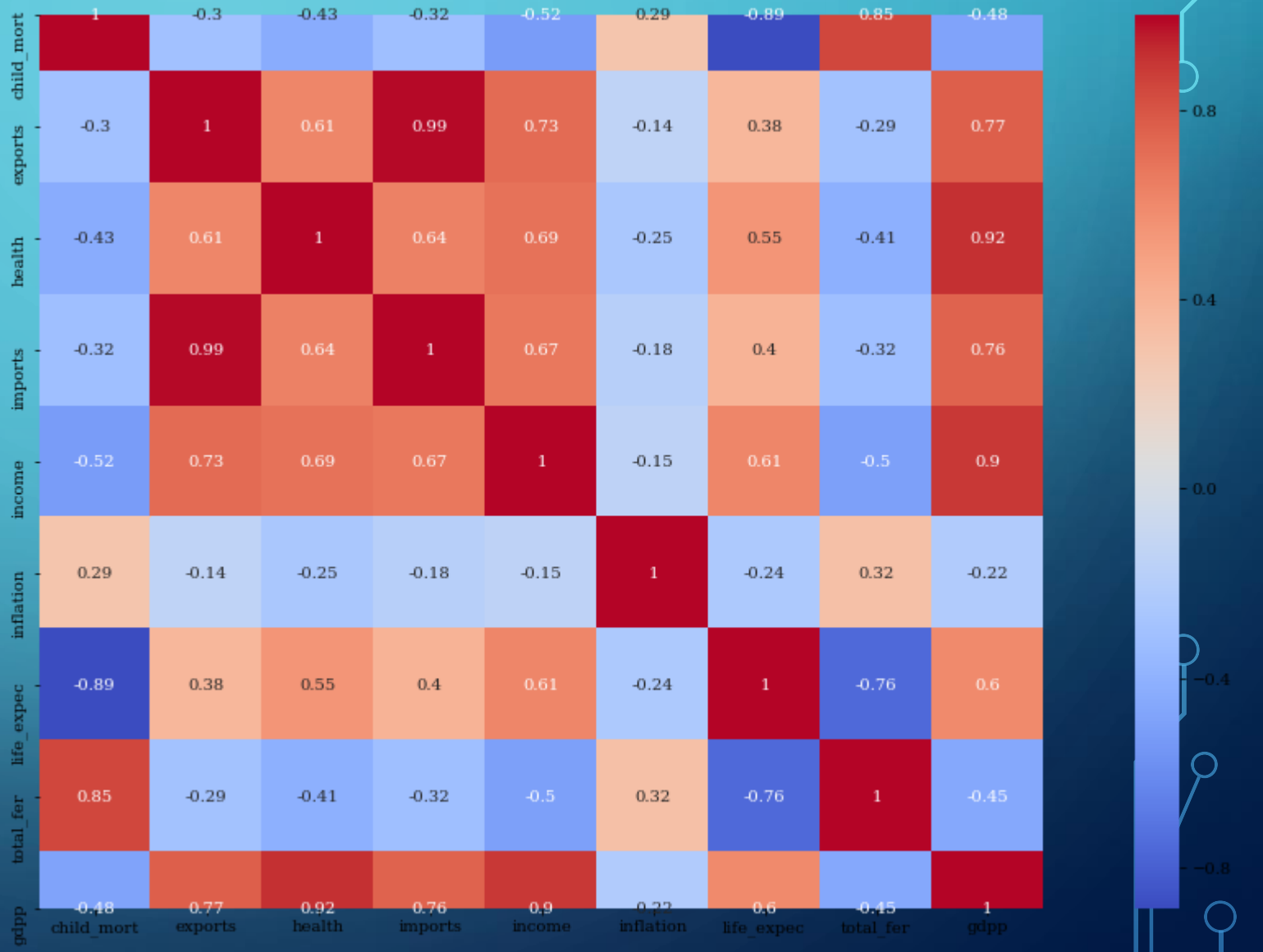
2) GDPP – List of countries with calculated Total GDP divided by the total population in ascending order.

3) Income – List of countries which are having net income per person is very low in ascending order.

HEATMAP

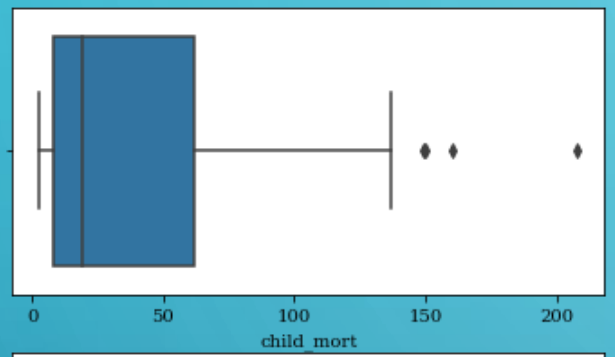
Looking at the heatmap we can see between import and exports (0.99), between health and gdpp(0.92), between income and gdpp(0.9) are highly (+vely) correlated with each other.

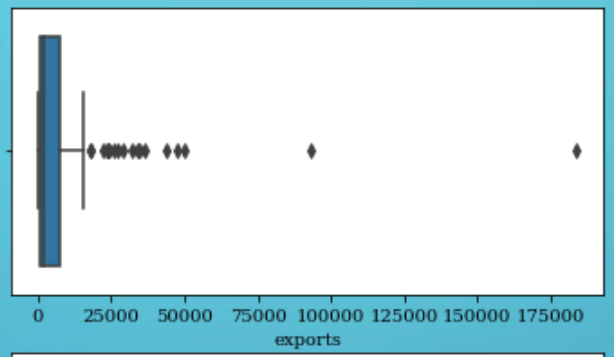
There are highly negatively correlated pair also like between child_mort and life_expec(-0.89) and between total_fertility and life_expec (-0.76).

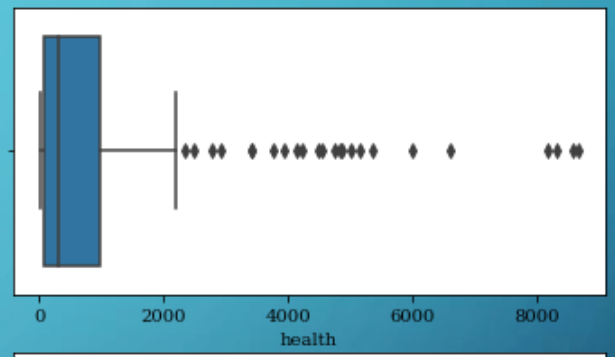


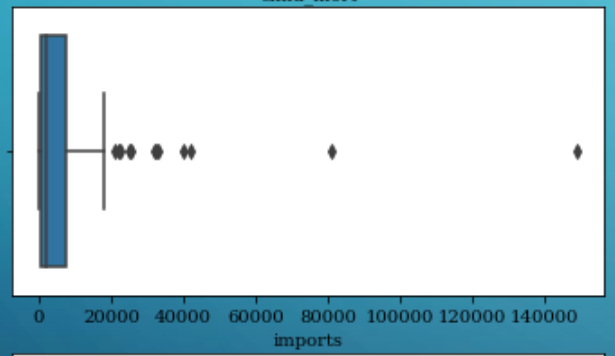
OUTLIER ANALYSIS

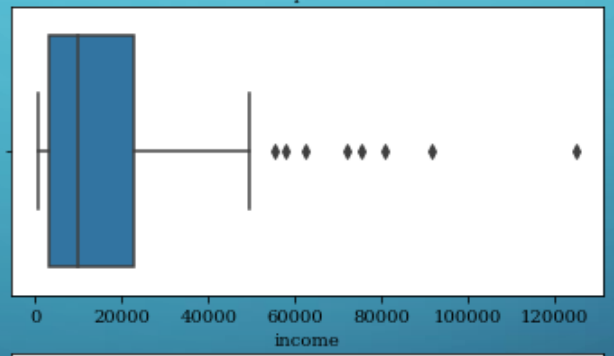
Outliers are present in almost all the columns and GDPP and Health are having more outliers present as compared to others.

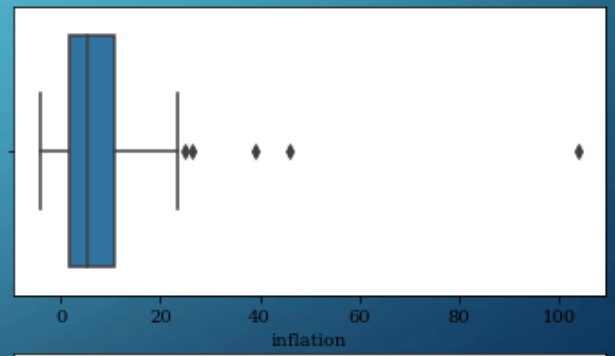


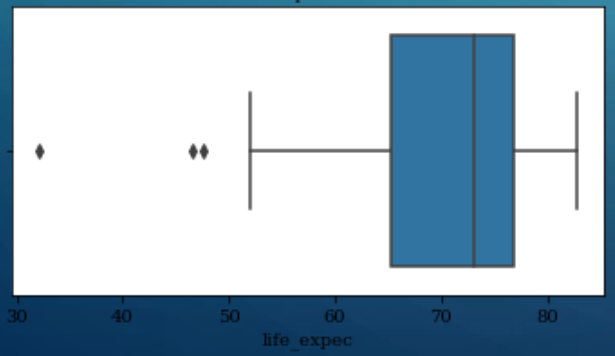


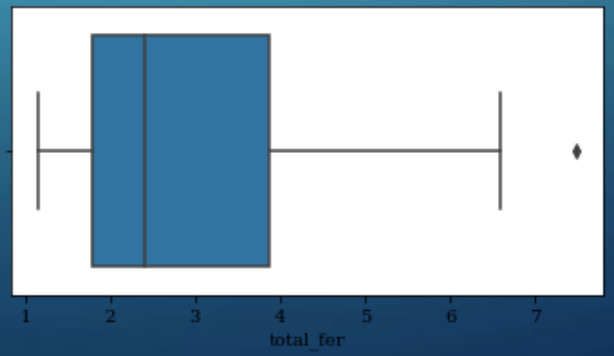


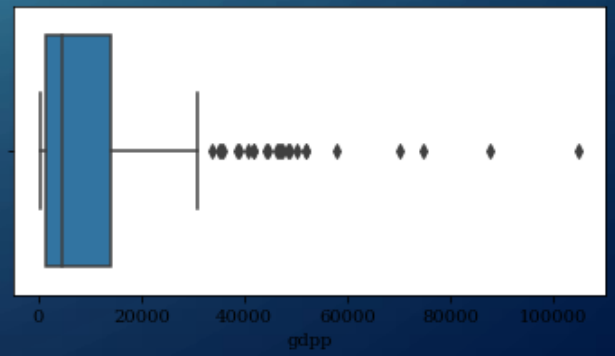






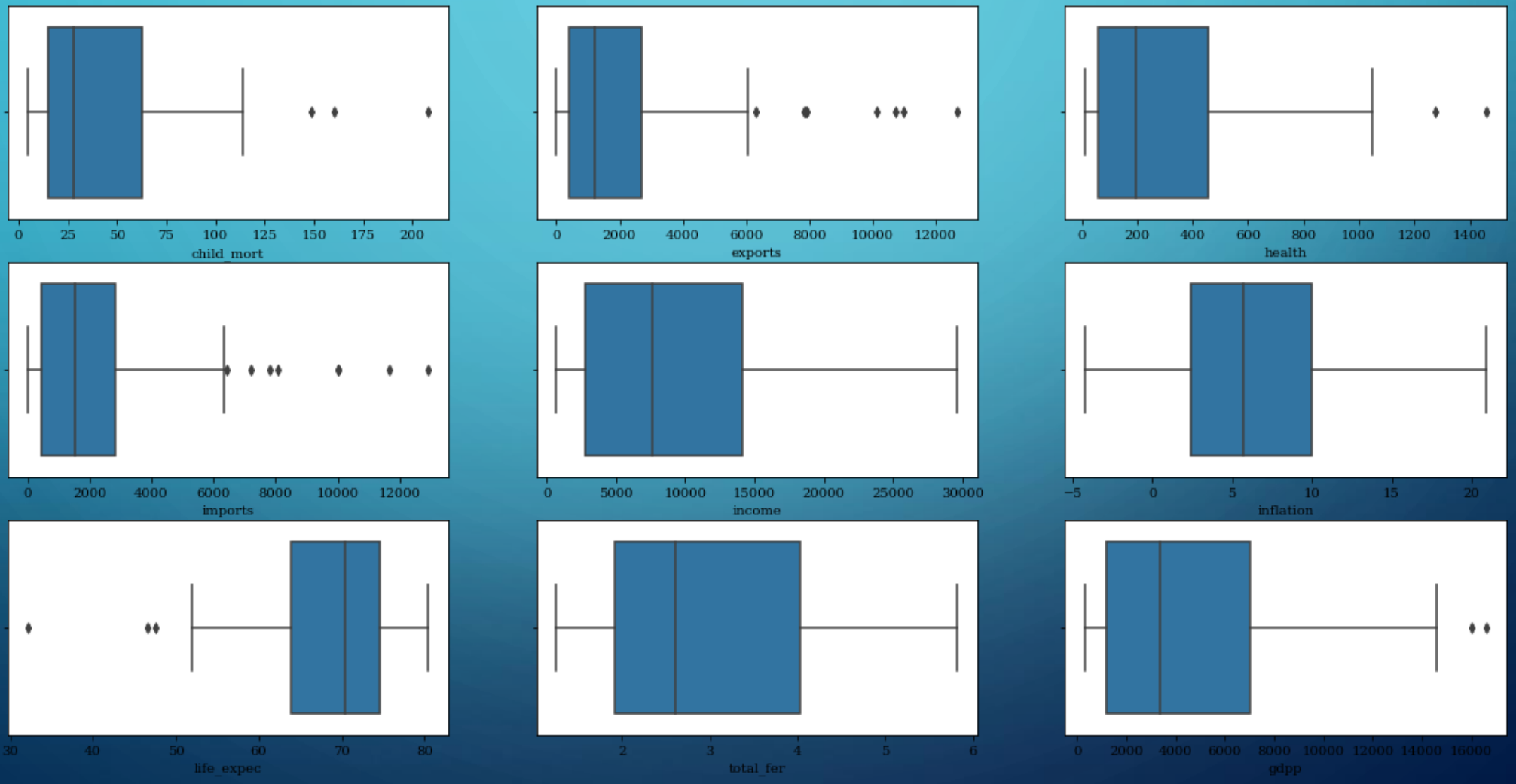




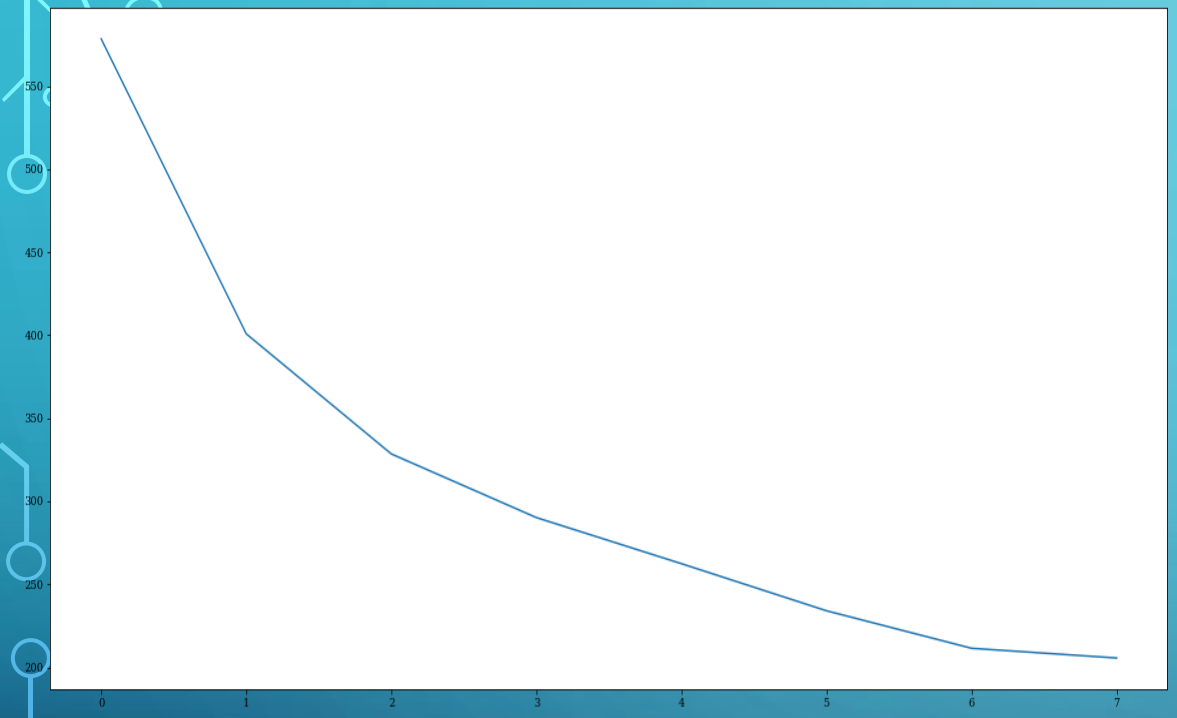


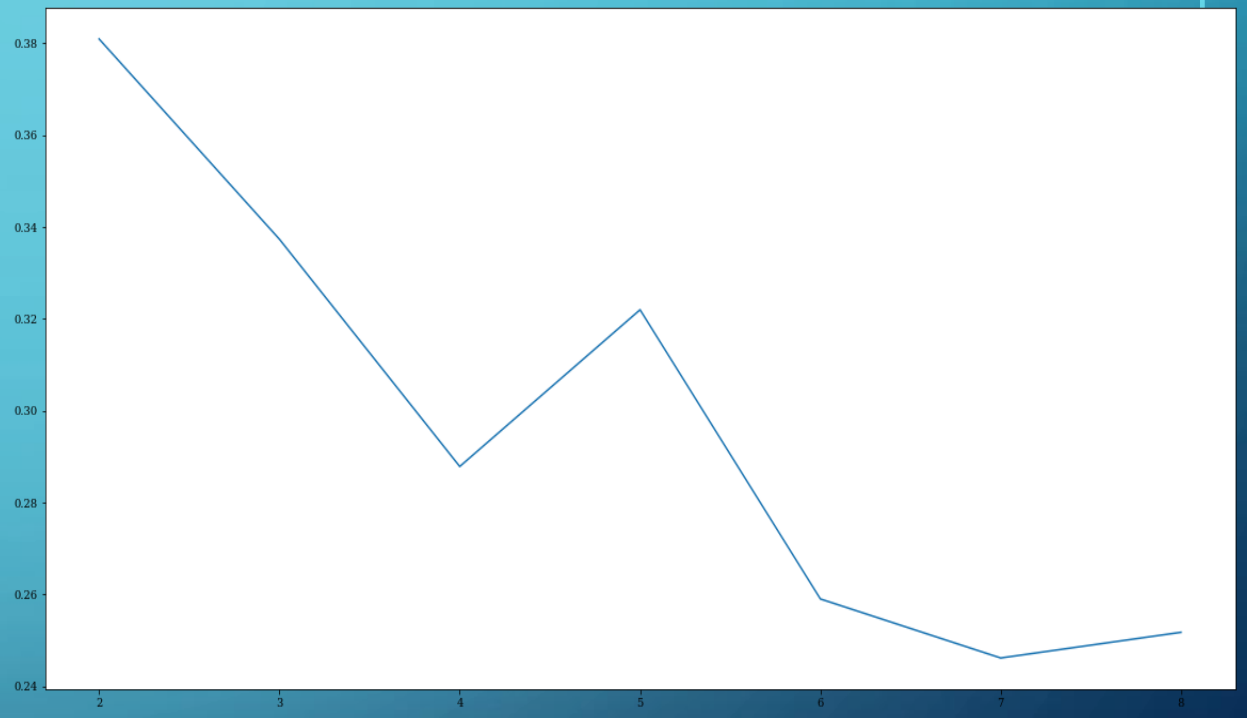
OUTLIER TREATMENT

Due to less dataset , it's not good to remove all the outliers from the dataset. Removing outliers grater than 0.95 quartile range using IQR.



K – MEANS CLUSTERING





Sum of Squared Method – We see that elbow curve range is between 2 - 4. Assume K = 3.

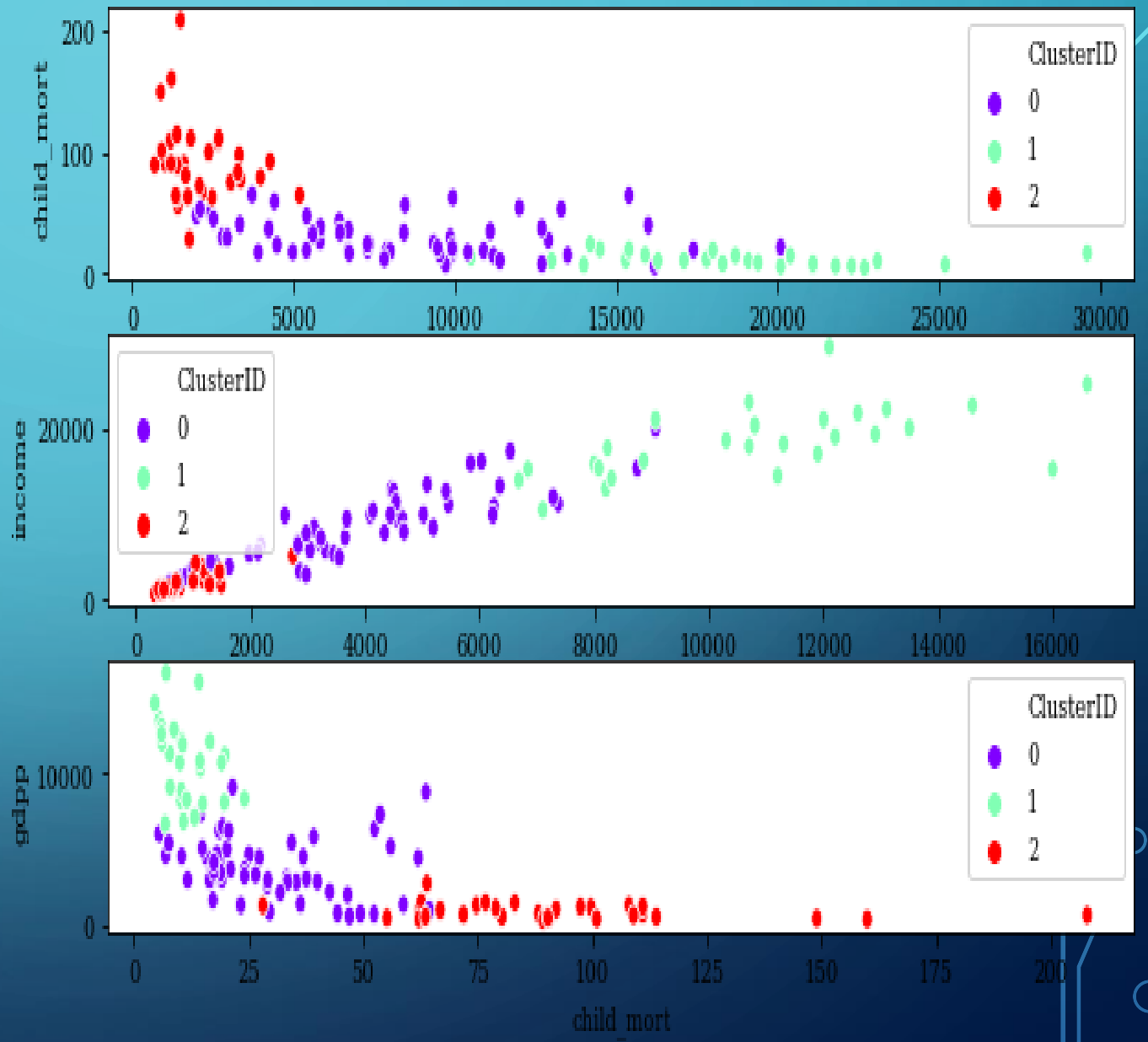
Silhouette Analysis – By observing 3 -5 are at highest peak and assuming K=3 for cluster analysis.

K-MEANS CLUSTERING

Child_mort – Cluster0 and cluster2 having high child mortality countries.

Income – Cluster0 and cluster2 having very low net income person.

GDPP – Cluster0 having very less calculated Total GDP divided by the total population.

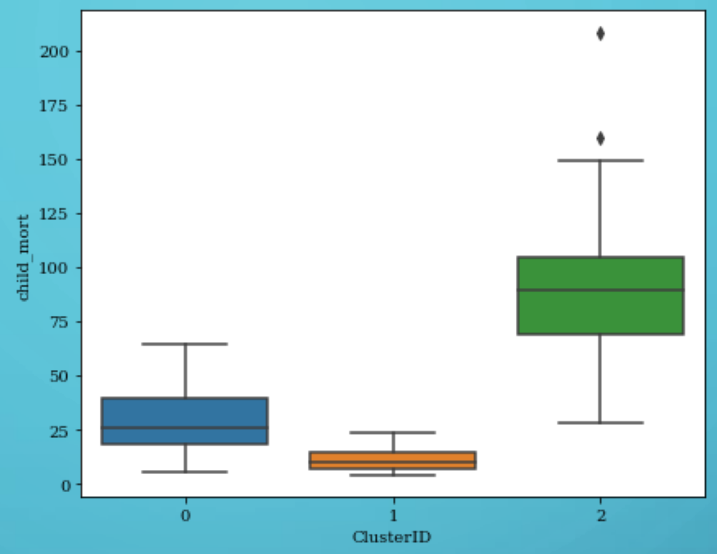


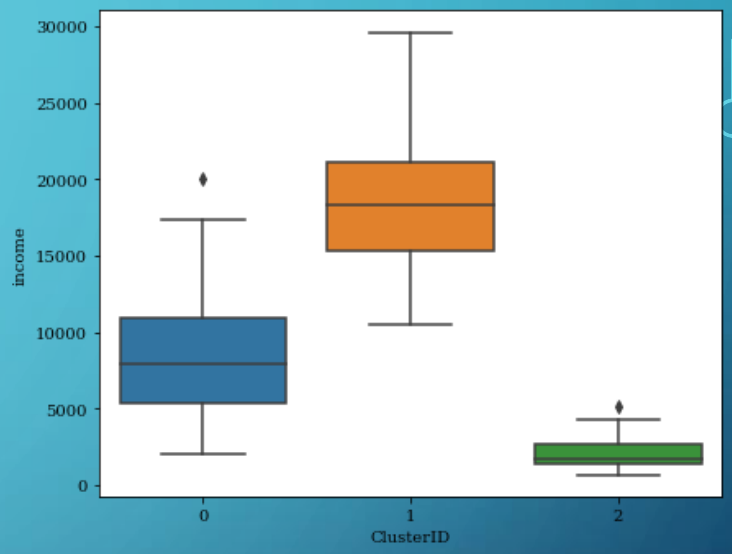
K-MEANS CLUSTERING

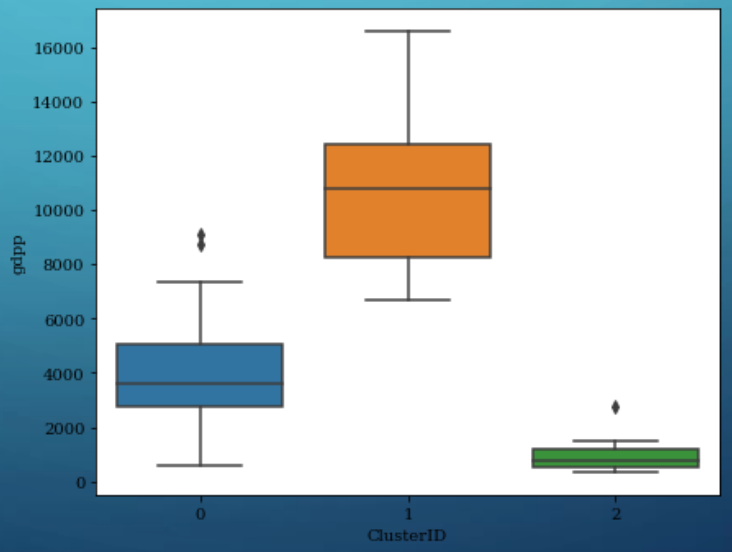
Cluster2 having high children mortality countries present and which needs to be aid for socio-economic factors. And Cluster1 having very less amount child mortality countries present.

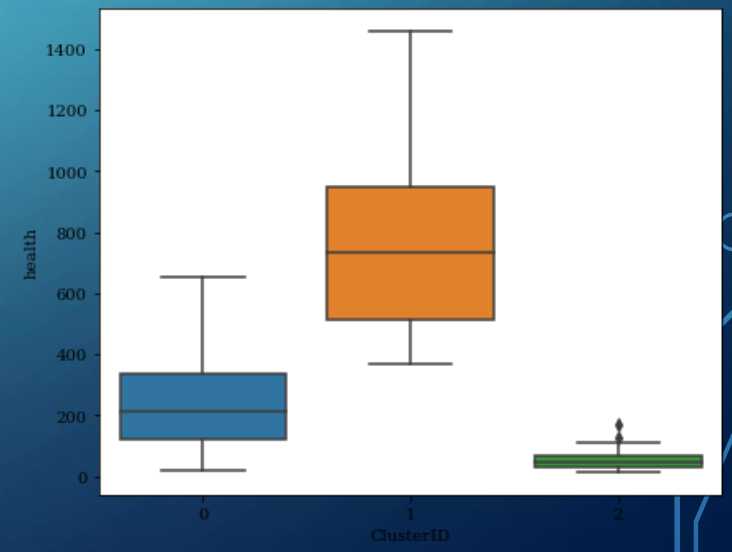
Cluster1 having countries with high income rate per person.

Cluster1 having high GDPP countries present.

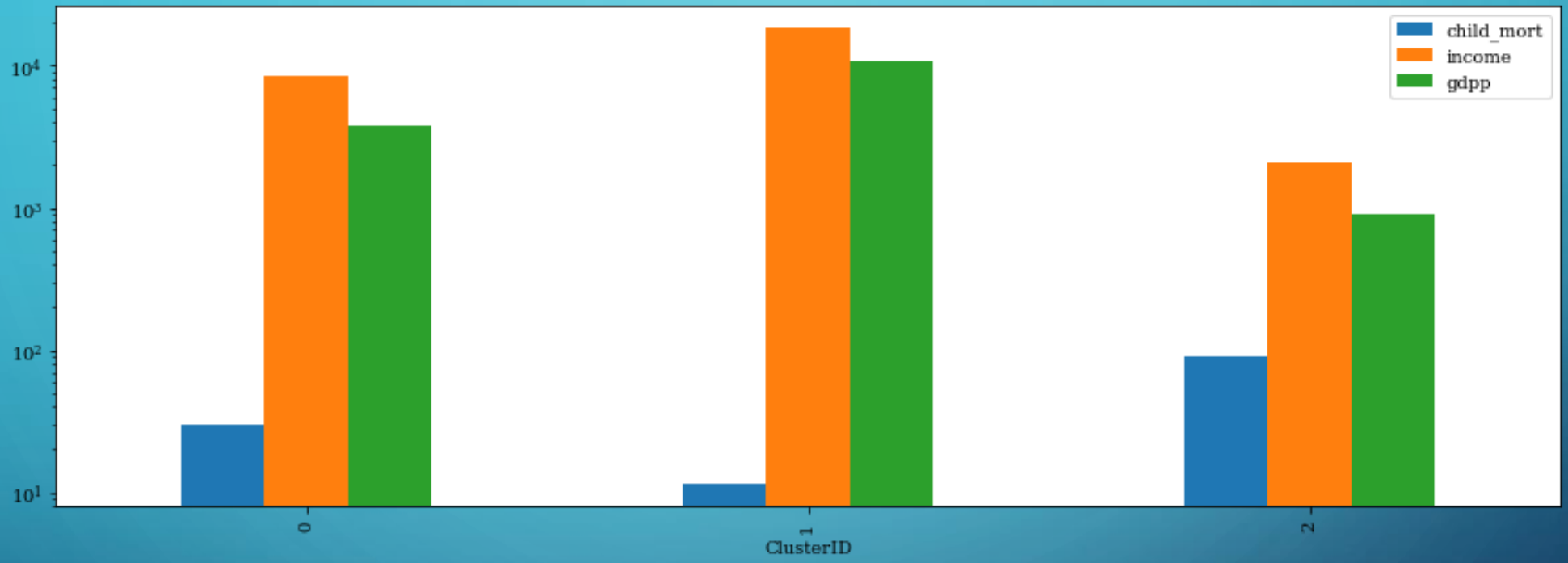








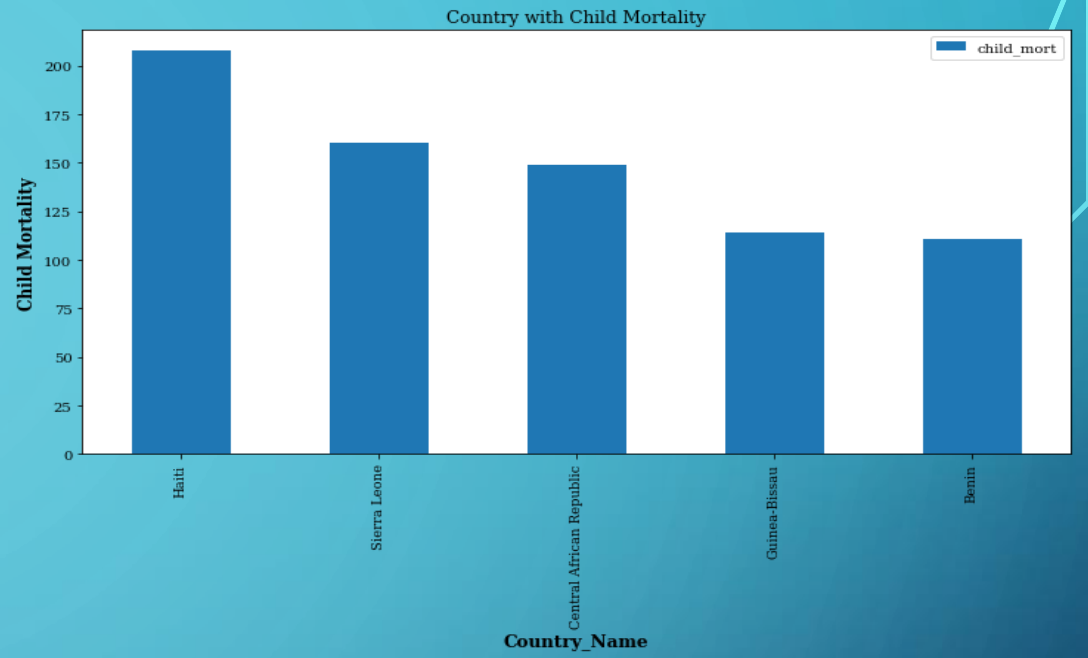
CLUSTER PROFILING IN K-MEANS

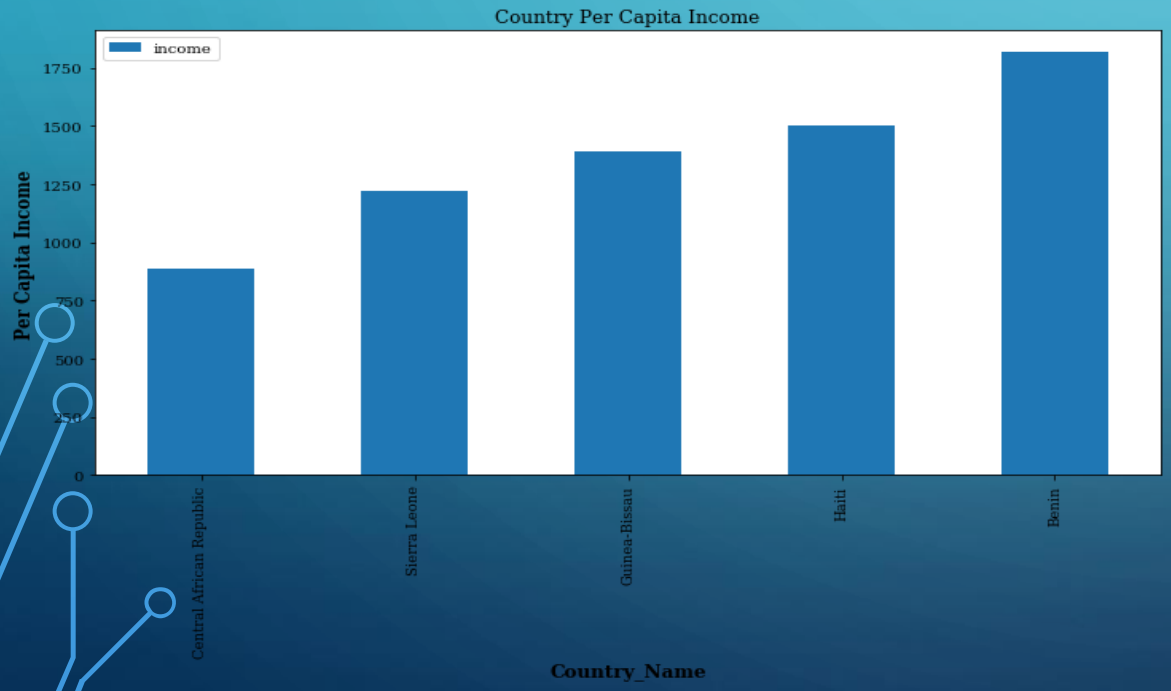


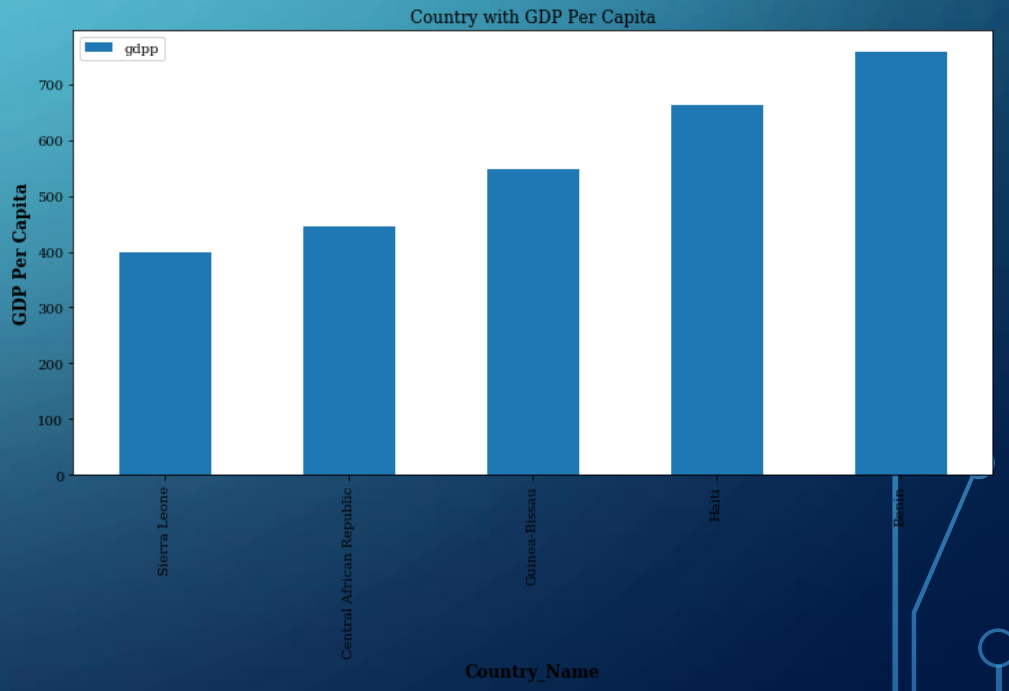
Cluster1 are having countries with very less children mortality, high income and high GDPP but we need to look on Cluster2 countries due to high children mortality and less income and less GDPP so need to be aid on for socio-economic factors.

CLUSTER PROFILING IN K-MEANS

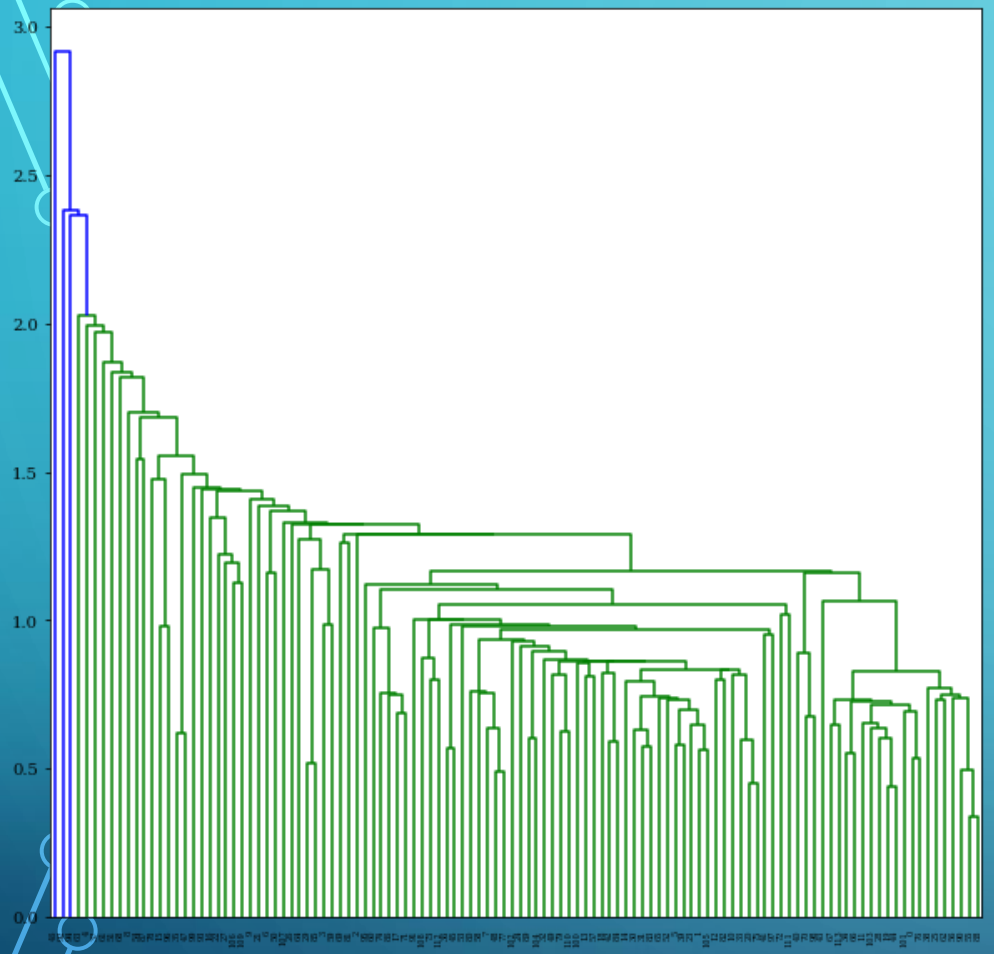
Countries with high children mortality and low income and low GDPP countries names.



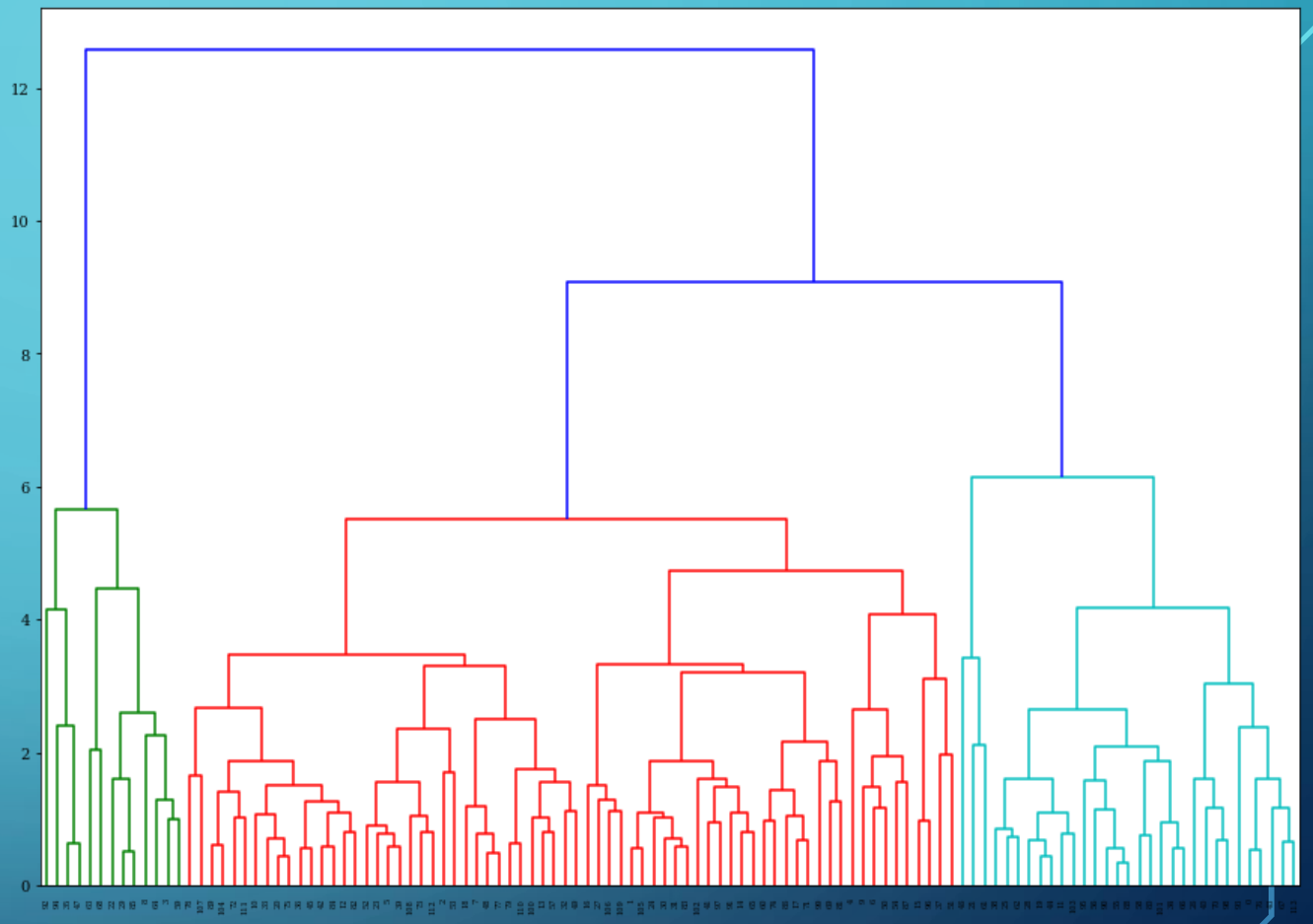




HIERARCHICAL CLUSTERING



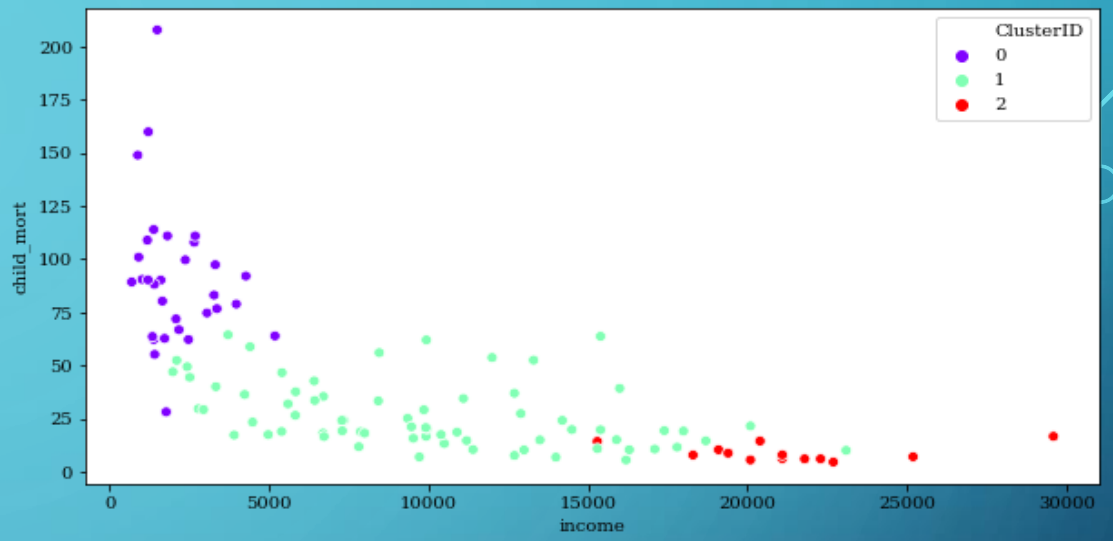
Single Linkage Hierarchical Method

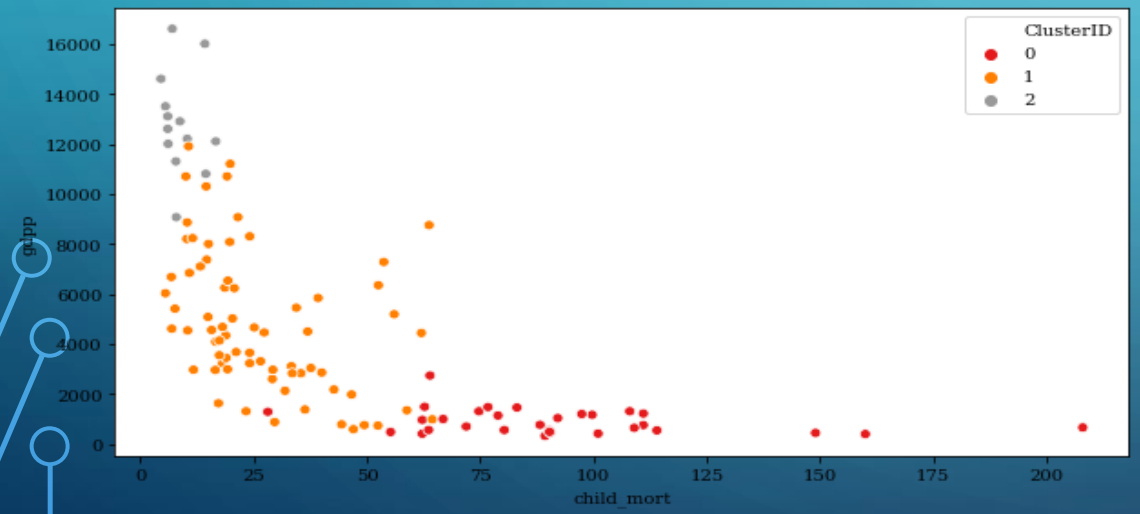


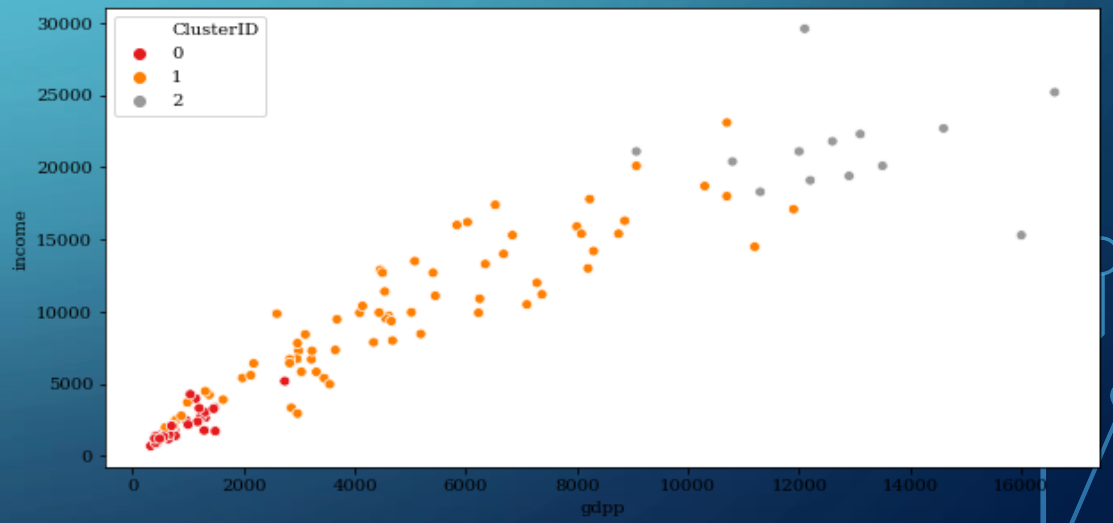
Complete Linkage Hierarchical clustering as single linkage clustering is not clear. By looking at dendograms taking n-cluster = 4.

HIERARCHICAL CLUSTERING

Scatter Plot with children mortality , income and GDPP values for cluster0 , and 2.





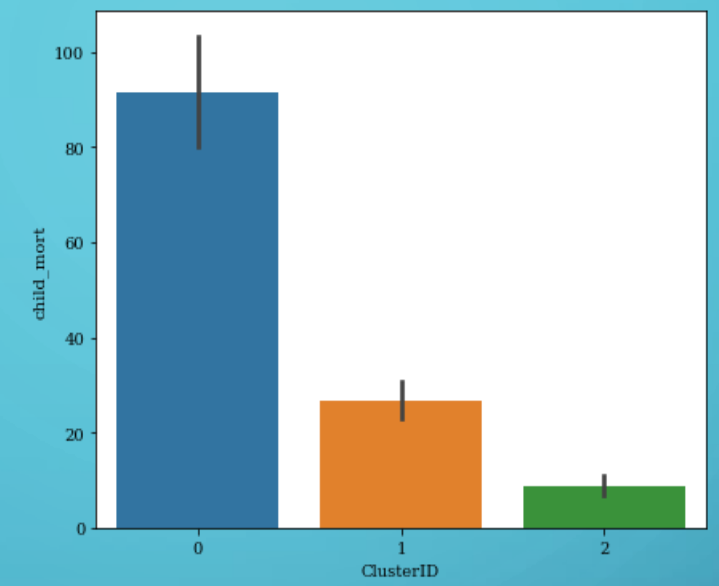


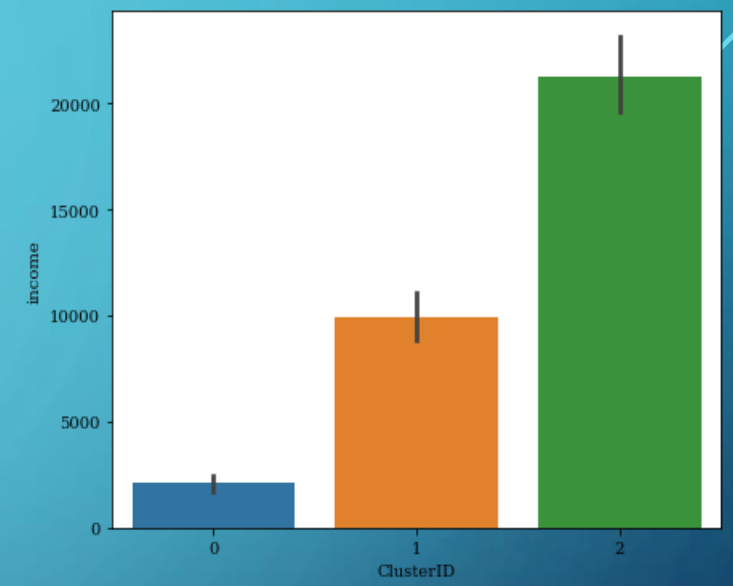
HIERARCHICAL CLUSTERING

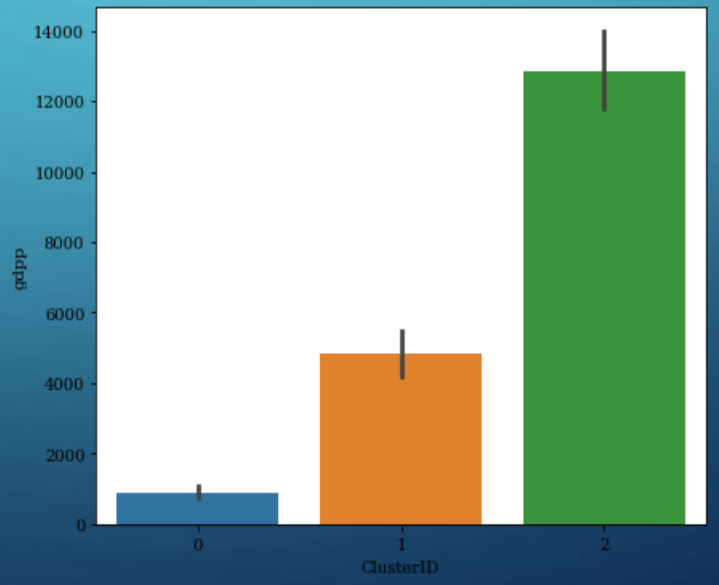
Children Mortality is high in cluster0 countries.

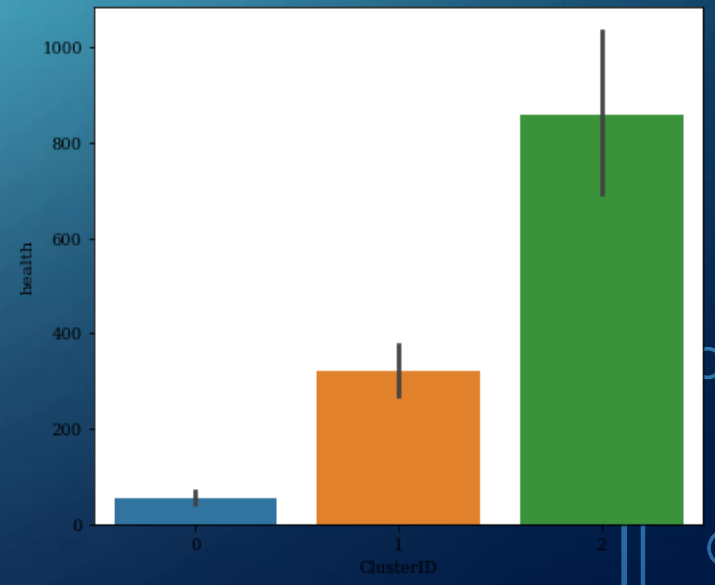
Income is high in cluster2 countries.

GDPP is high in cluster2 countries.

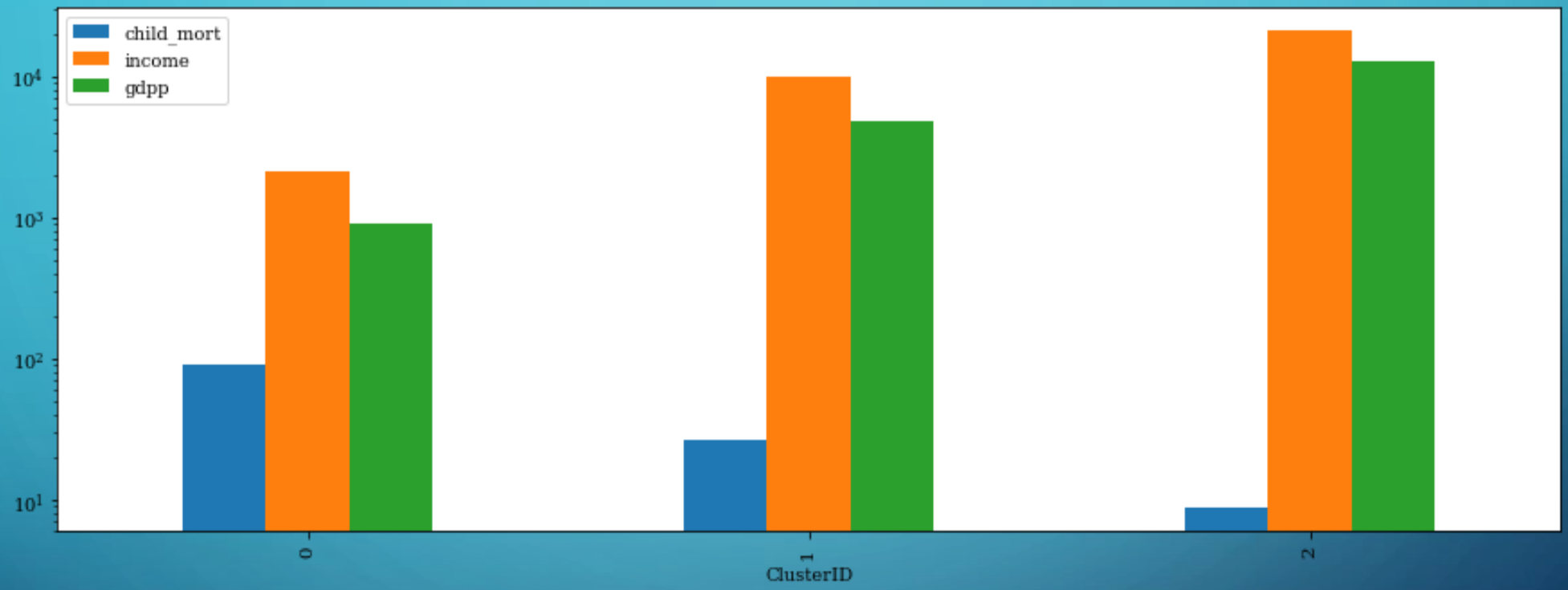








CLUSTER PROFILING ON HIERARCHICAL

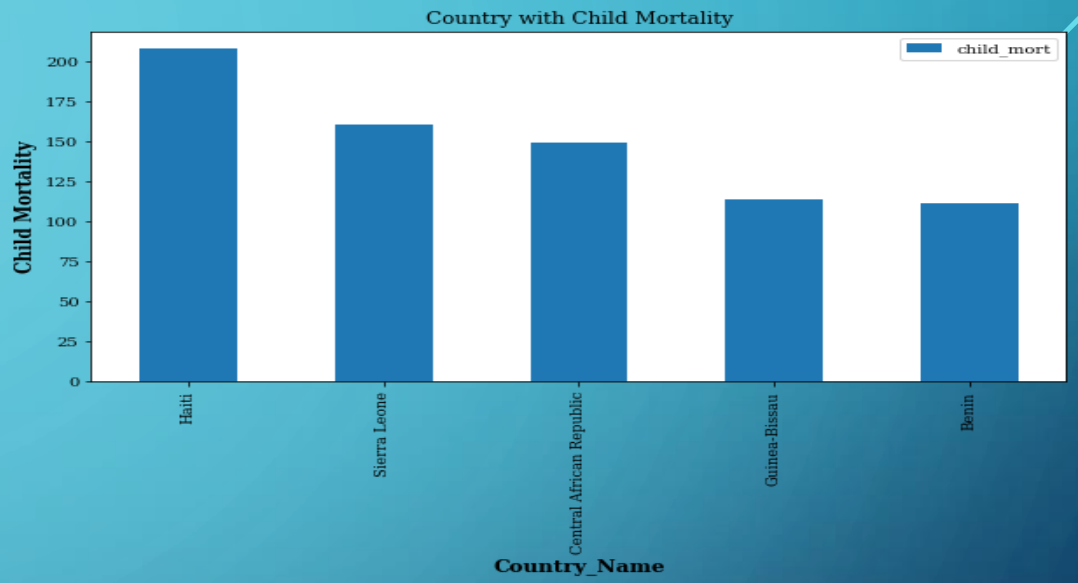


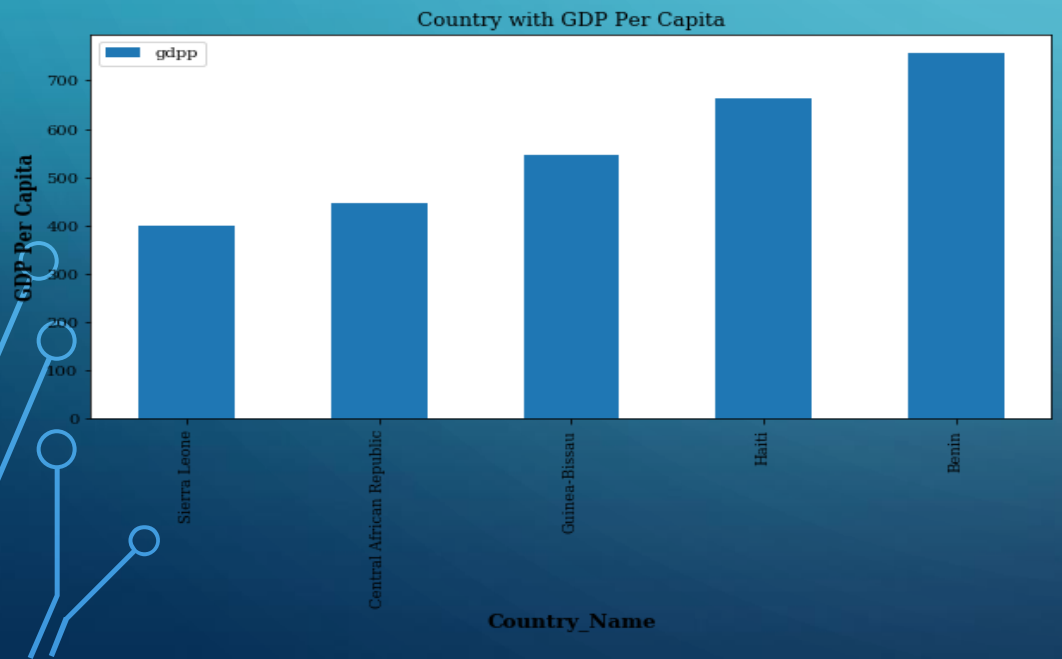
Cluster0 having more no of children mortality as compared to others.

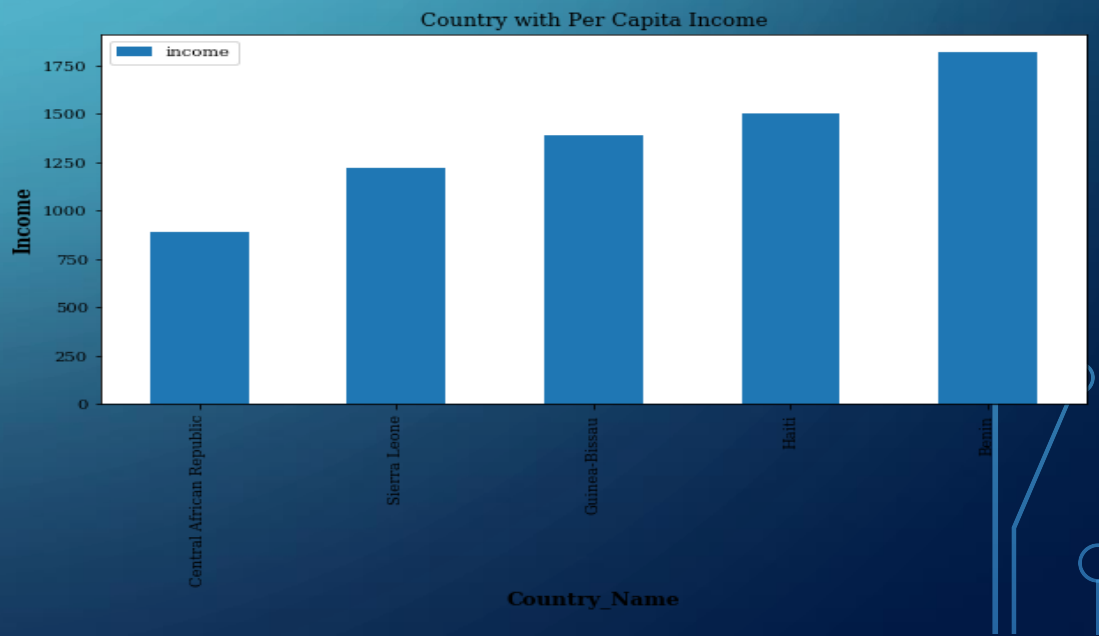
Cluster2 are having more no income and GDPP countries present and very less amount of children mortality.

CLUSTER PROFILING ON HIERARCHICAL

Countries with high children mortality and low income and low GDPP countries names.







SUMMARY

By performing K – Means and Hierarchical clustering , got the same list of countries which requires aid by considering socio-economic factors into consideration.

```
#Final Contry Names with K-Means
Final_country_by_KM.reset_index(drop=True).country
```

```
0                        Haiti
1                 Sierra Leone
2    Central African Republic
3                Guinea-Bissau
4                        Benin
Name: country, dtype: object
```

```
# Final countries Names with Hierarchical clustering
Final_country_by_HC.reset_index(drop=True).country
```

```
0                        Haiti
1                 Sierra Leone
2    Central African Republic
3                Guinea-Bissau
4                        Benin
Name: country, dtype: object
```