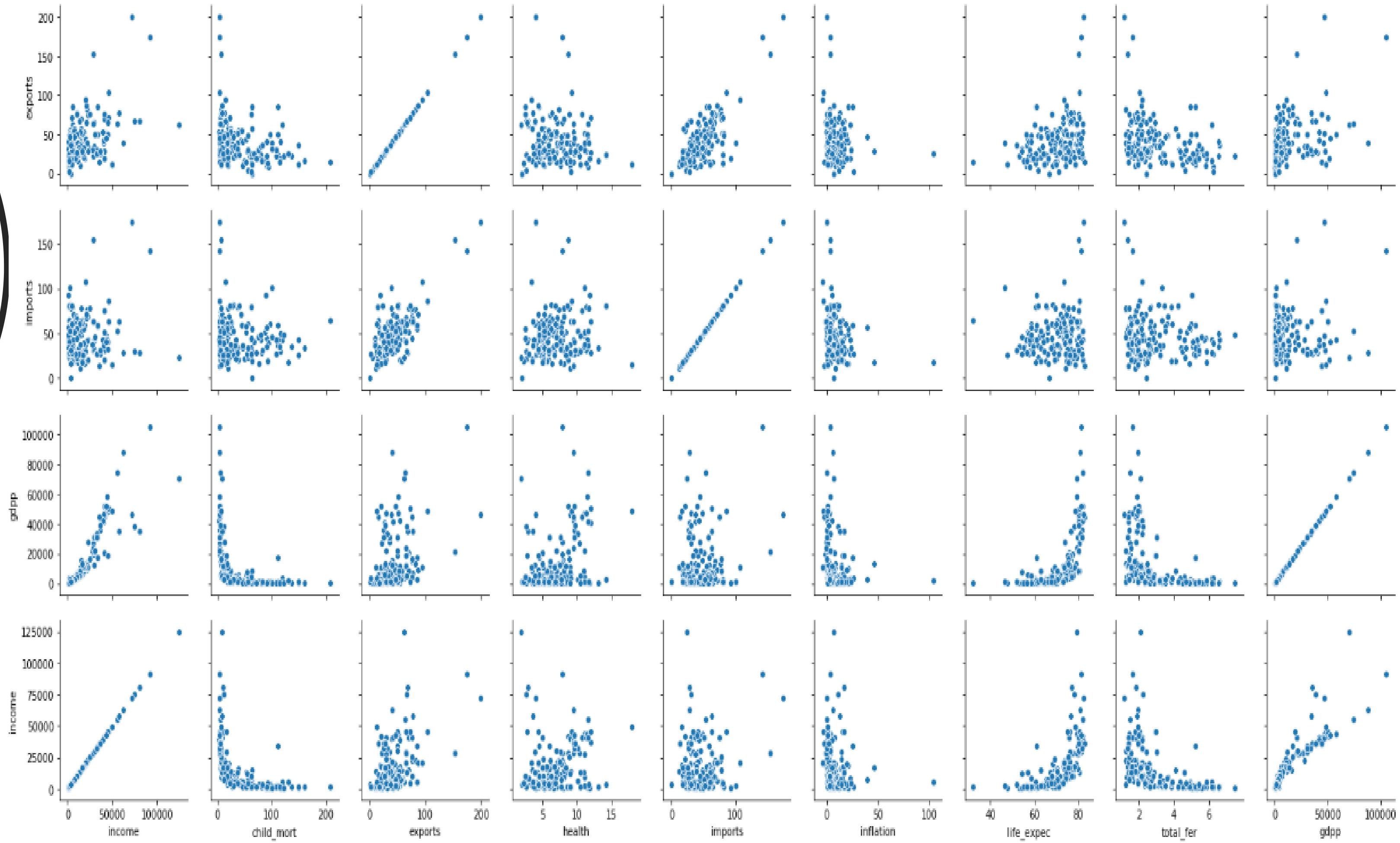
PCA and Clustering Assignment

- 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. It runs a lot of operational projects from time to time along with advocacy drives to raise awareness as well as for funding purposes.
- Objective is to recommend the countries that are in direst need of aid.

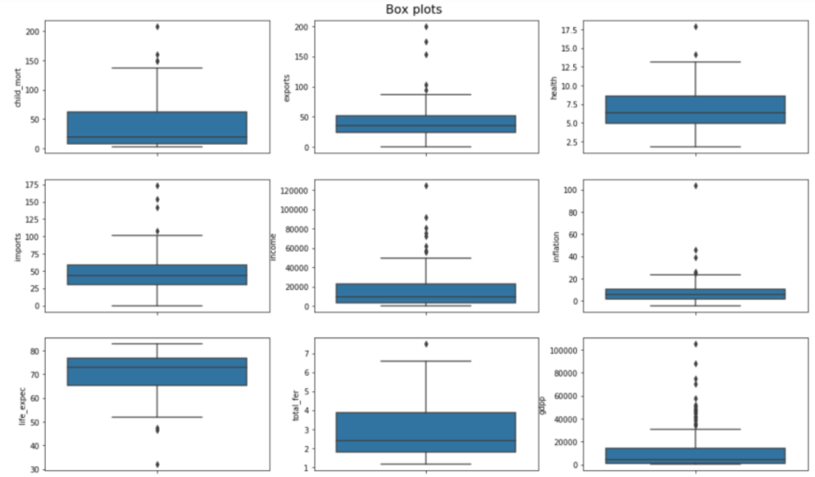
Approach

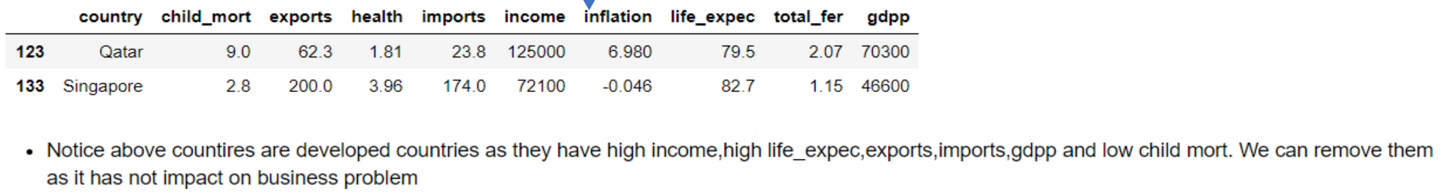
- Loading data and understanding the data by visualizing
- Removed outliers
- Scaled the data
- Applied principal components analysis
- Removed outlier after PCA
- Applied K-Means clustering
- Applied Hierarchical clustering
- Derived the list of countries

Loading and understanding the data



- Understanding the spread of the data to check for outliers
- Opted to remove outliers from income, exports, imports, gdpp variables as these variables will be higher for developed countries. So, it has not impact on business problem

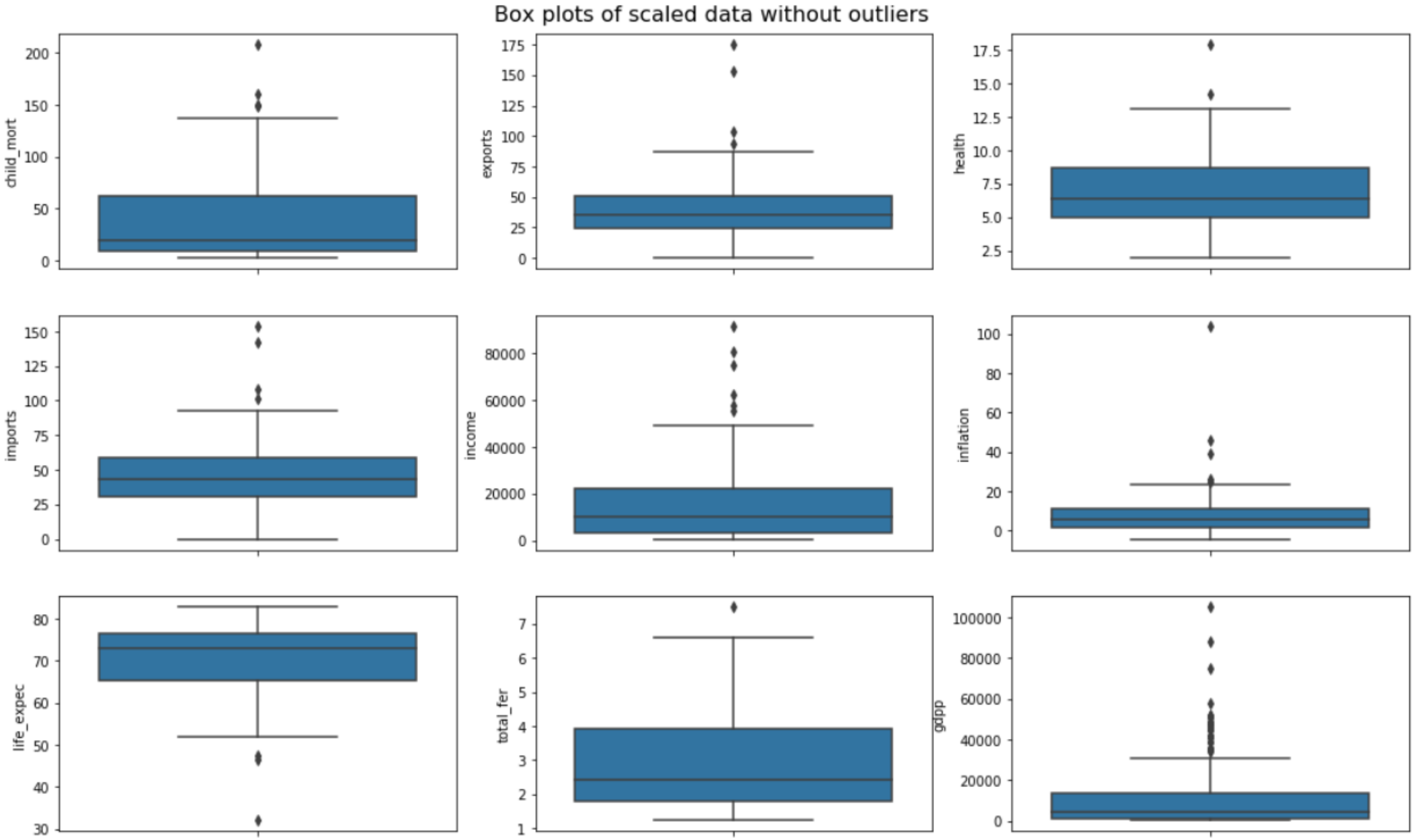




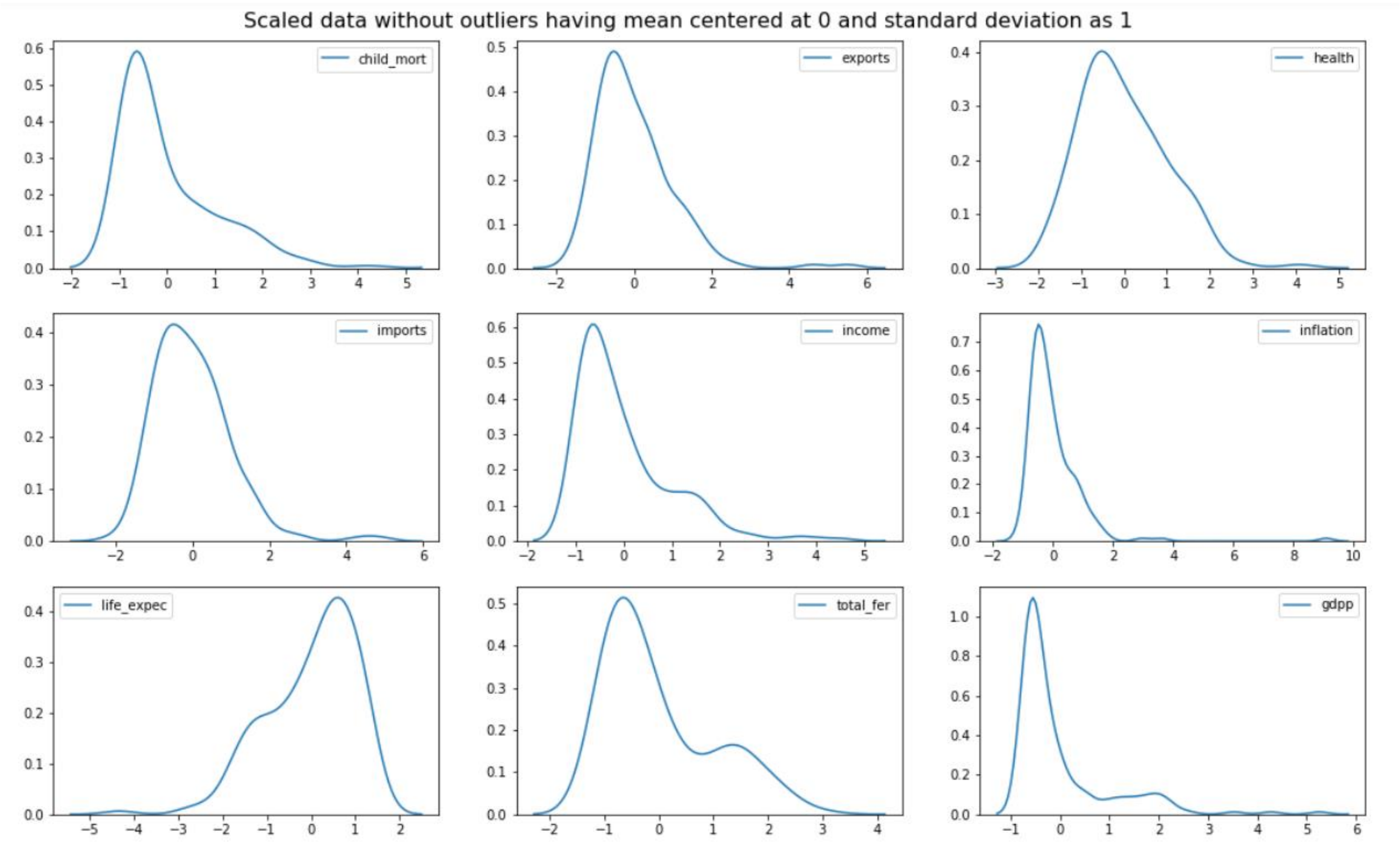
	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
123	Qatar	9.0	62.3	1.81	23.8	125000	6.980	79.5	2.07	70300
133	Singapore	2.8	200.0	3.96	174.0	72100	-0.046	82.7	1.15	46600

- Notice above countires are developed countries as they have high income,high life_expec,exports,imports,gdpp and low child mort. We can remove them as it has not impact on business problem

Data spread after removing outliers

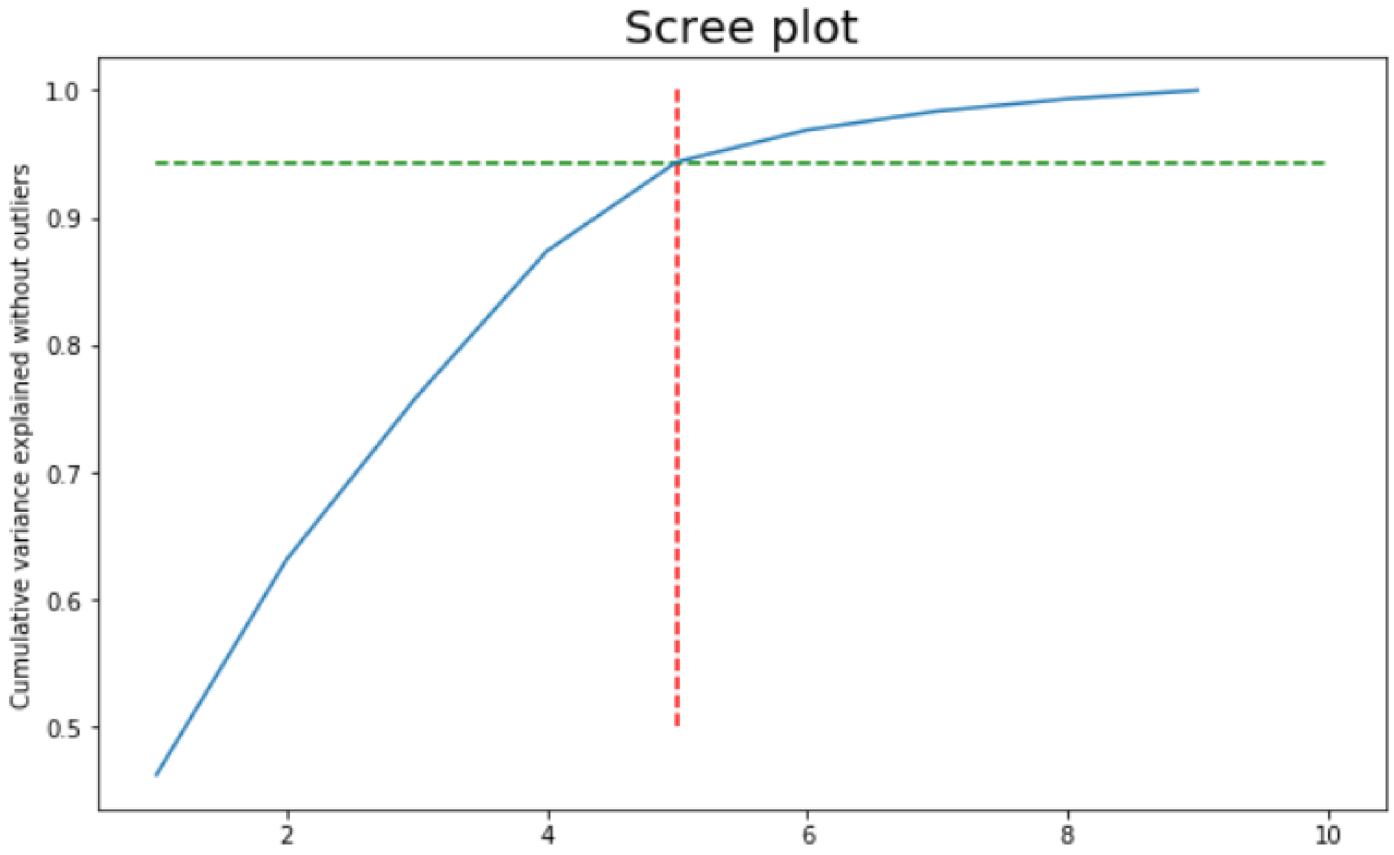


Data spread after scaling



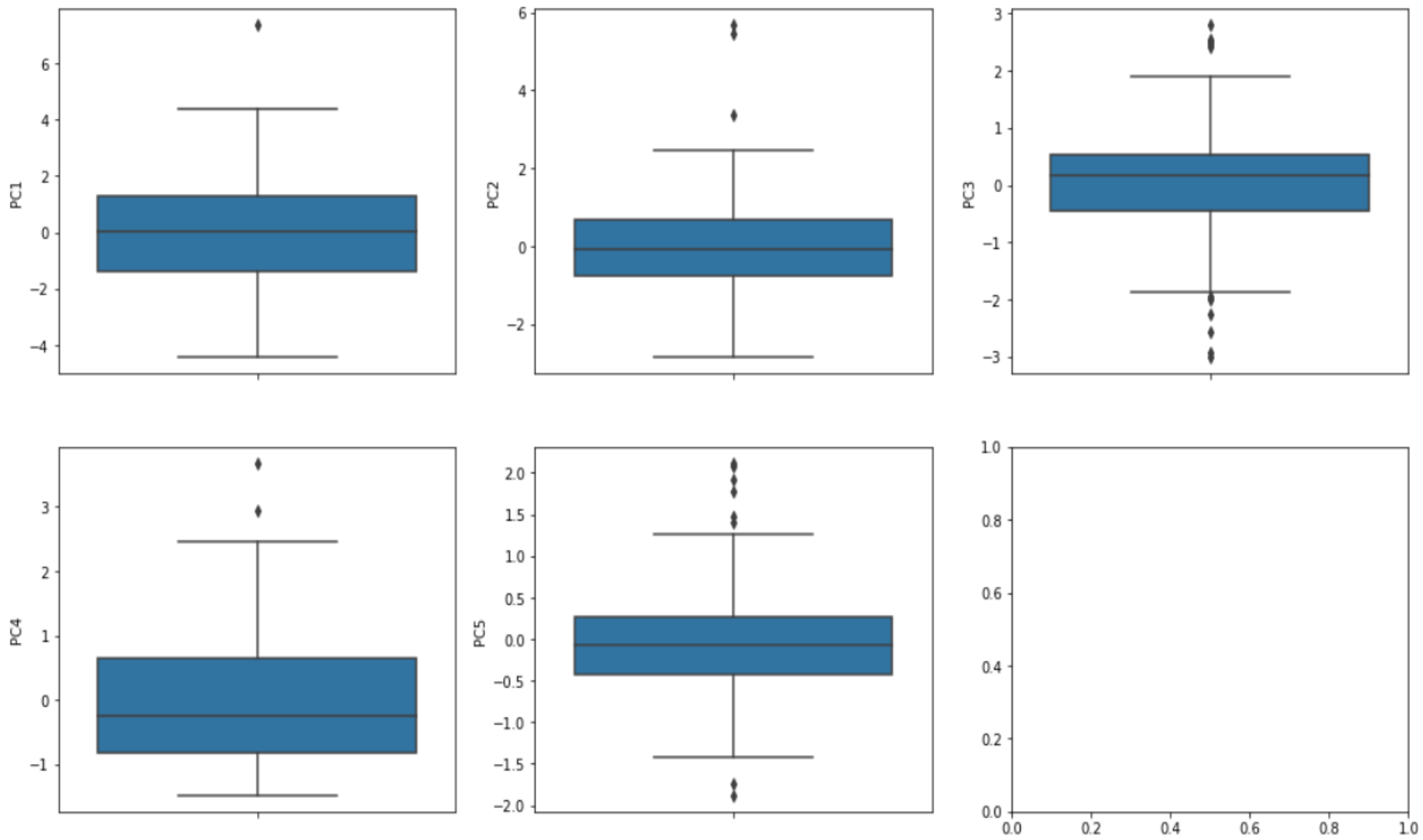
Scaled data without outliers having mean centered at 0 and standard deviation as 1

Scree plot to determine optimum number of principal components



5 principal components explain about 94.5% of variance

Box plot of the principal components after removing outliers

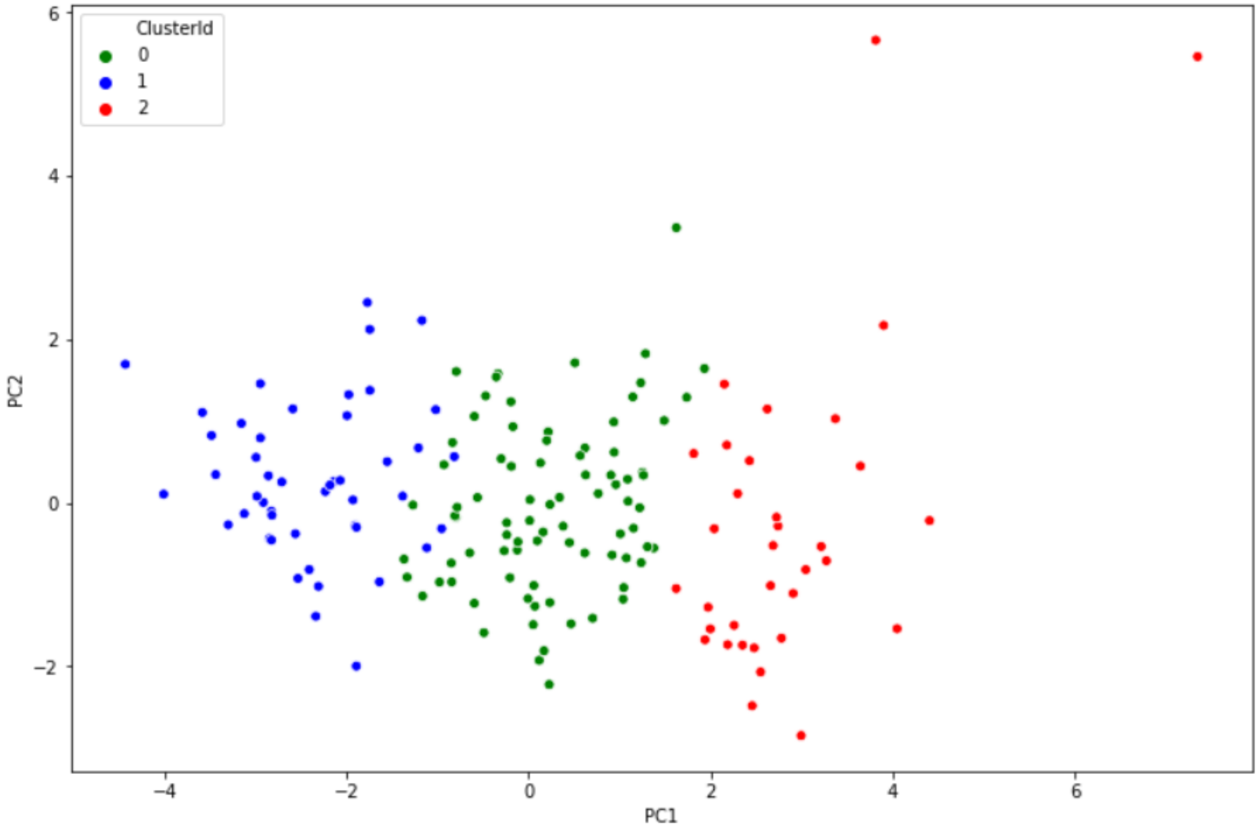


- 5 principal components are formed after performing PCA.

Hopkins test

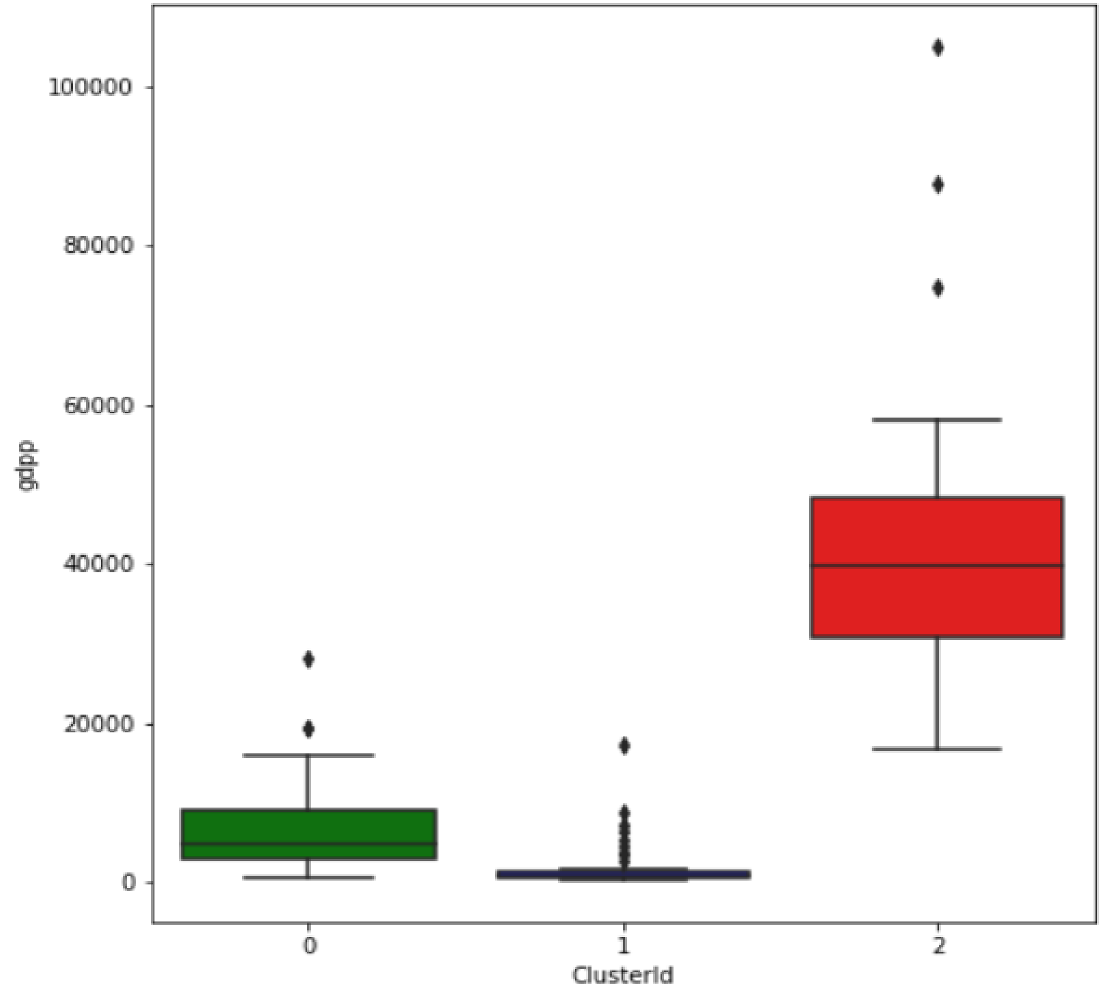
- Hopkins test gave a result ranging 75 to 86.
- This show data can be used for clustering.

Cluster plot after performing K-Means clustering

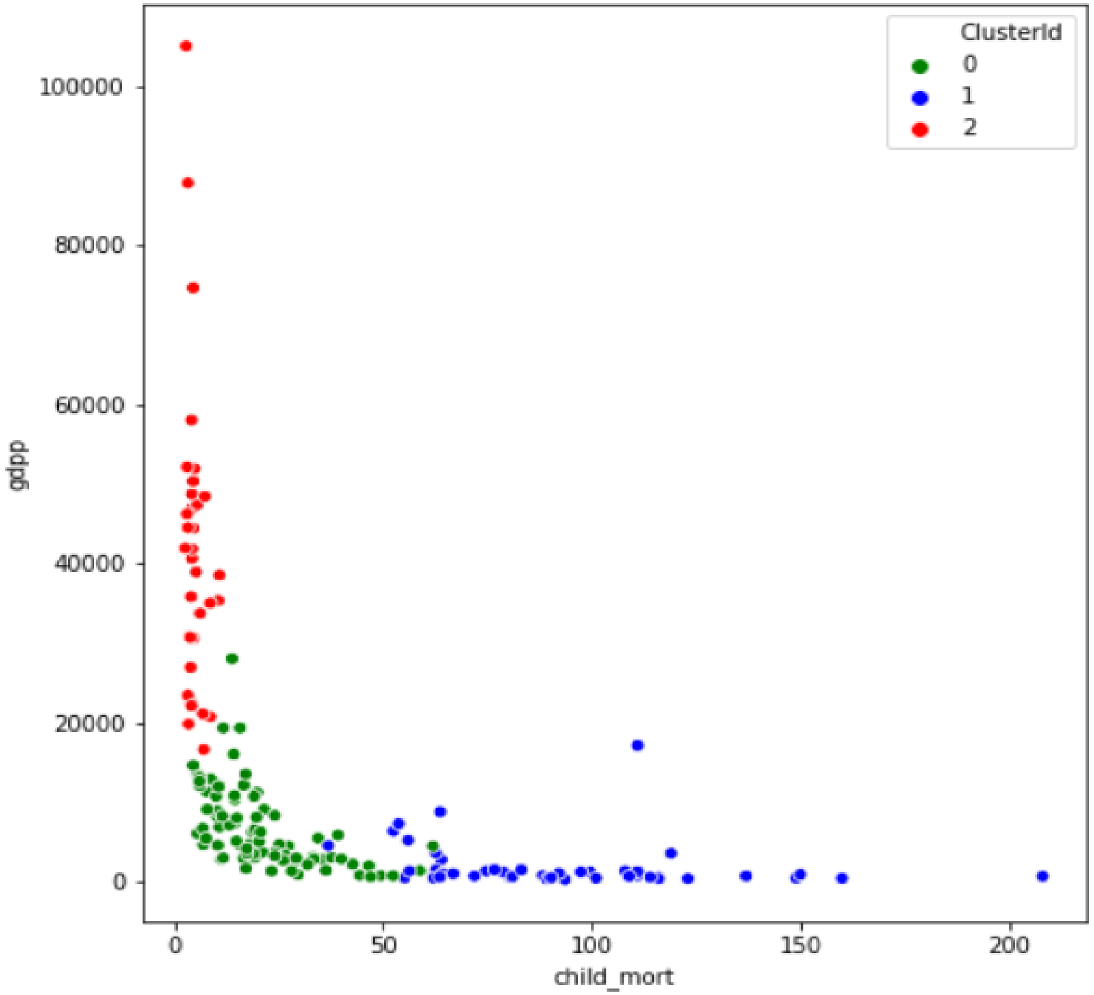


- 3 Cluster are formed
- Formed cluster are tight with good cohesion

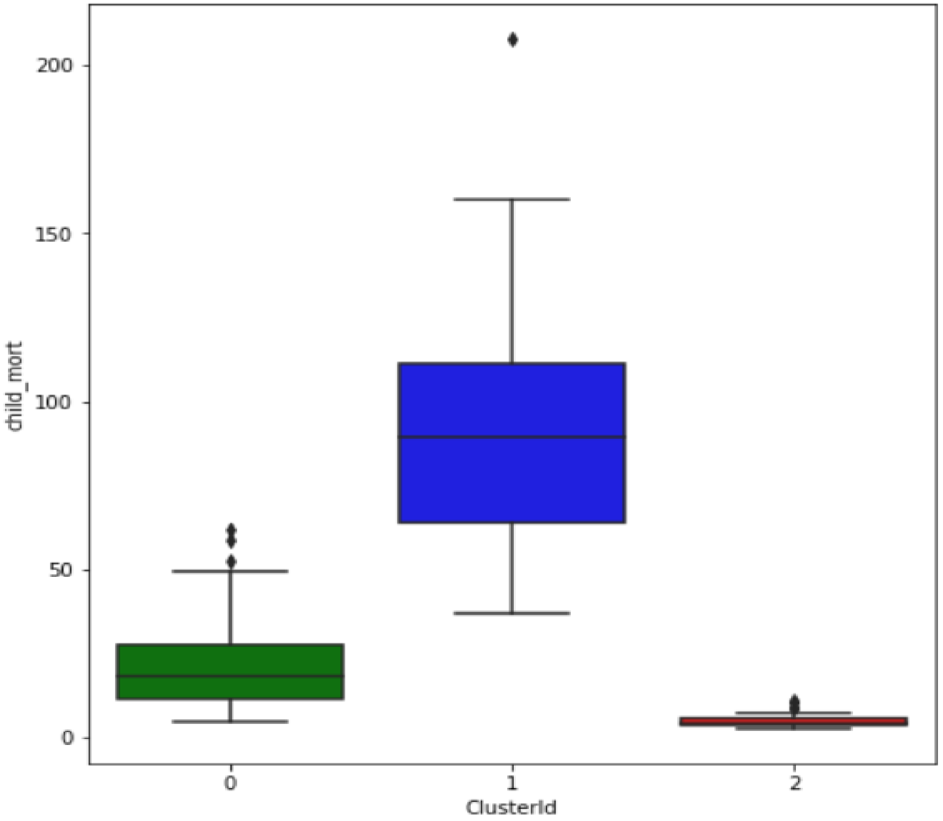
Box plot between cluster Id and GDP per capita



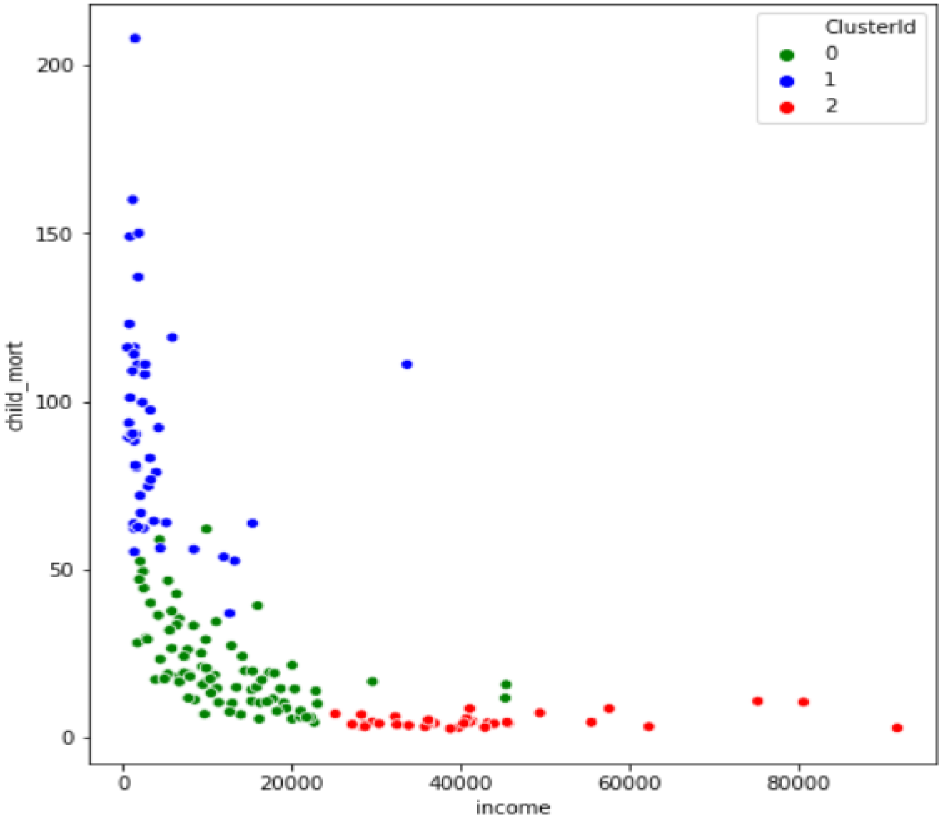
Box plot between child mortality rate and GDP per captia



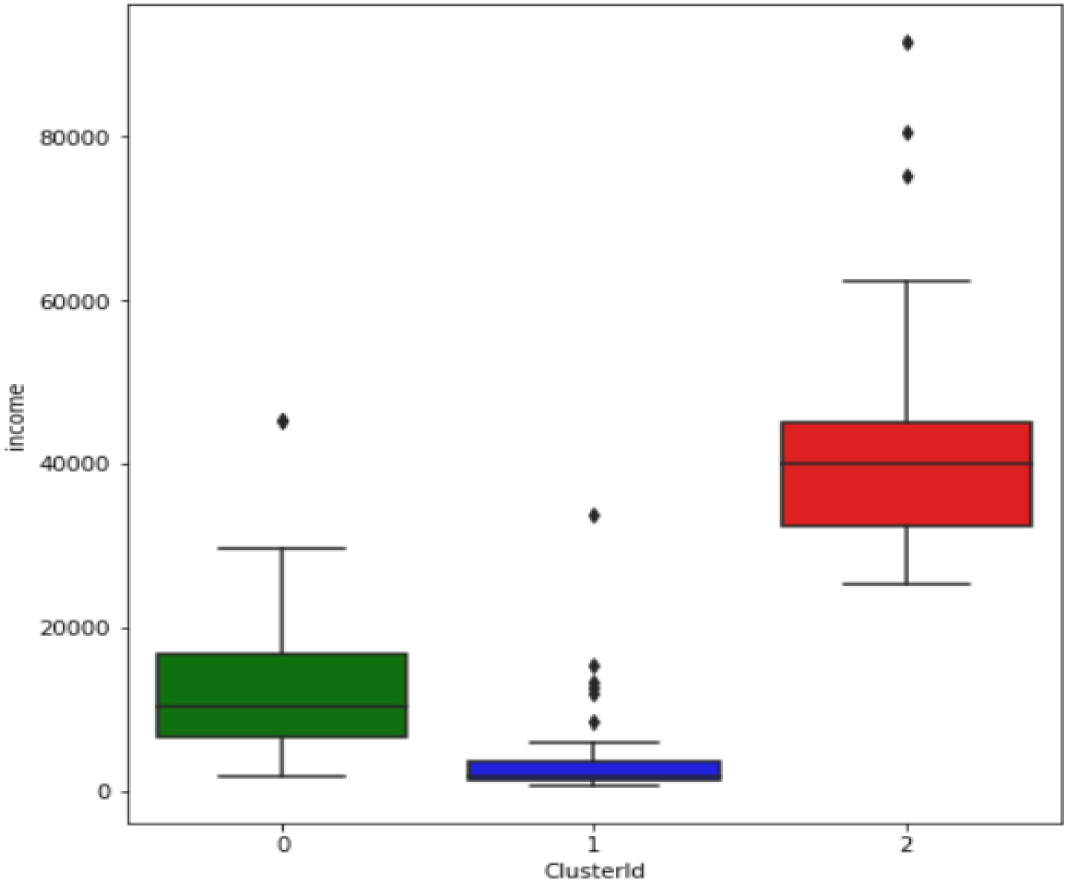
Box plot between cluster Id & child mortality rate



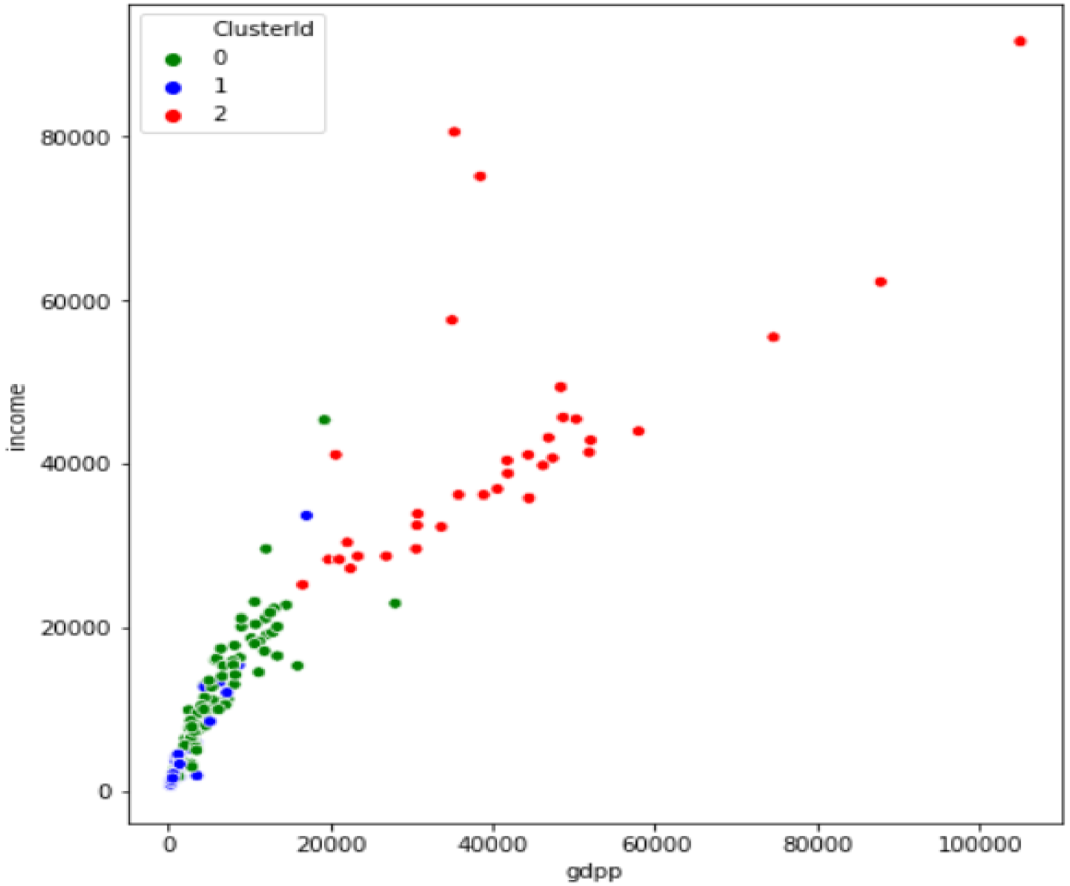
Scatter plot between income and child mortality rate



Box plot between cluster Id & income



Box plot between GDP per capita and income

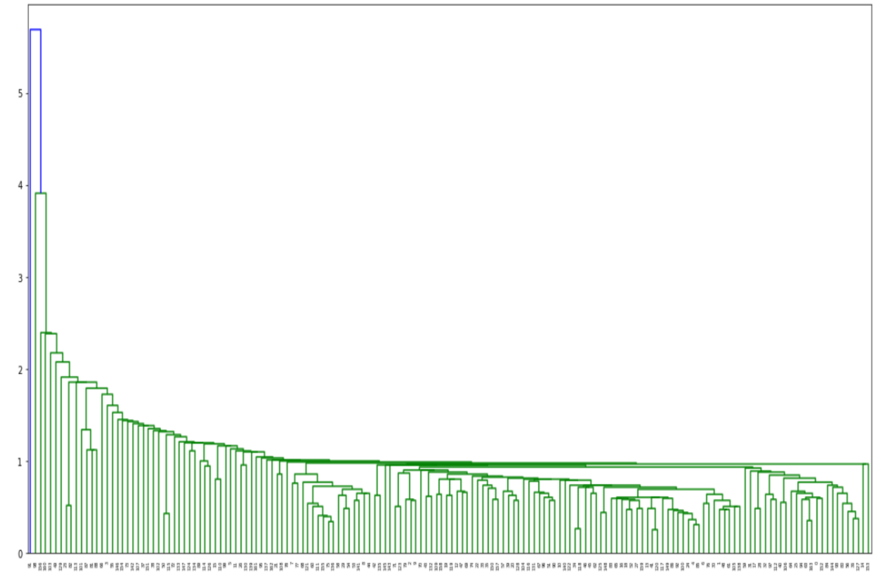


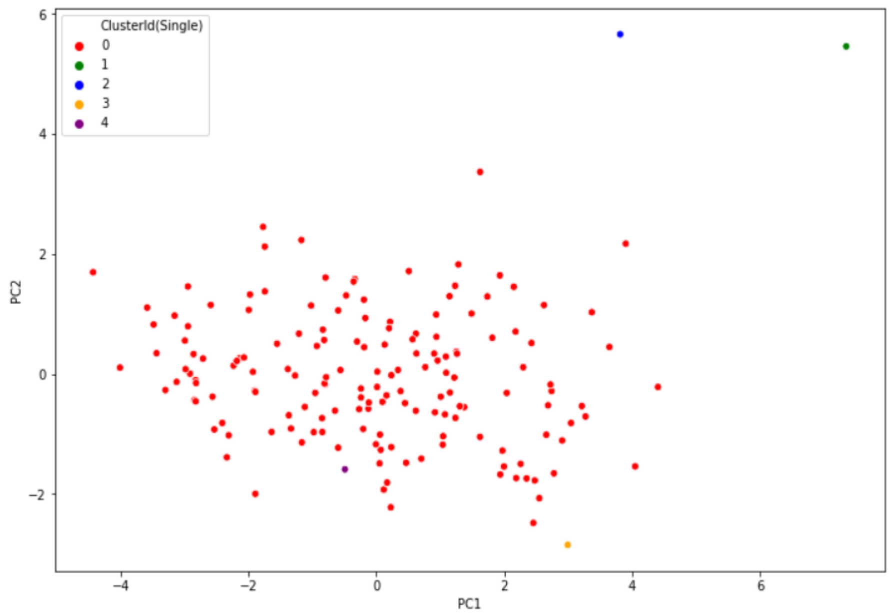
Conclusion from box plots

- Countries have cluster Id as 1(one) are in need of aid
  - As they have low income
  - Low GDP per captia
  - High child mortality rate

Hierarchical clustering

- Single Linkage
- Dendrogram and cluster plot on data after applying single linkage clustering
    - Clusters are very loose hasn't produced desired results.

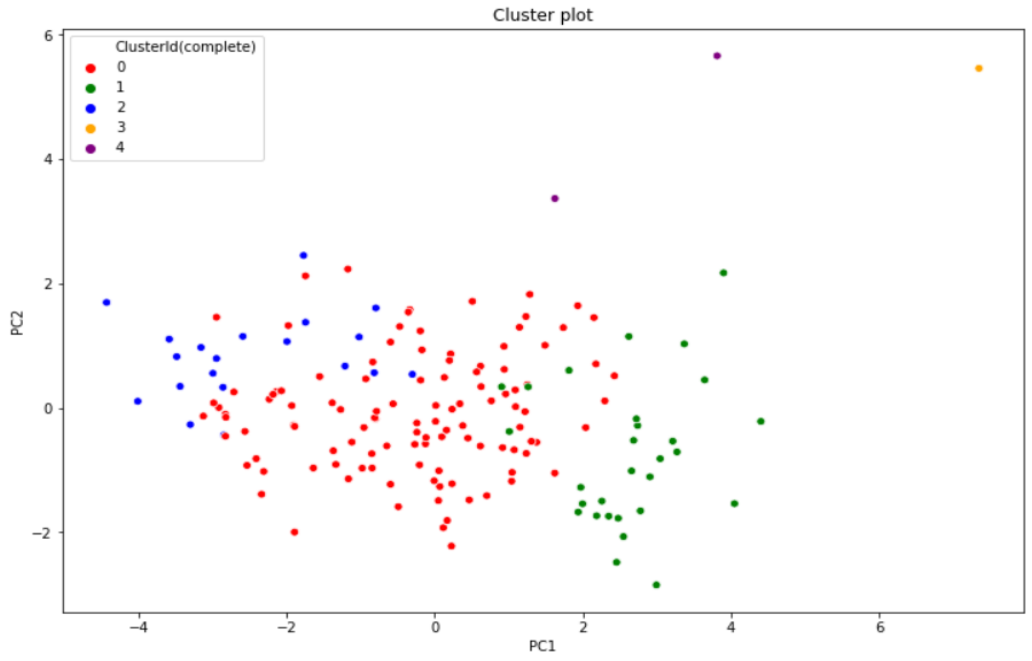




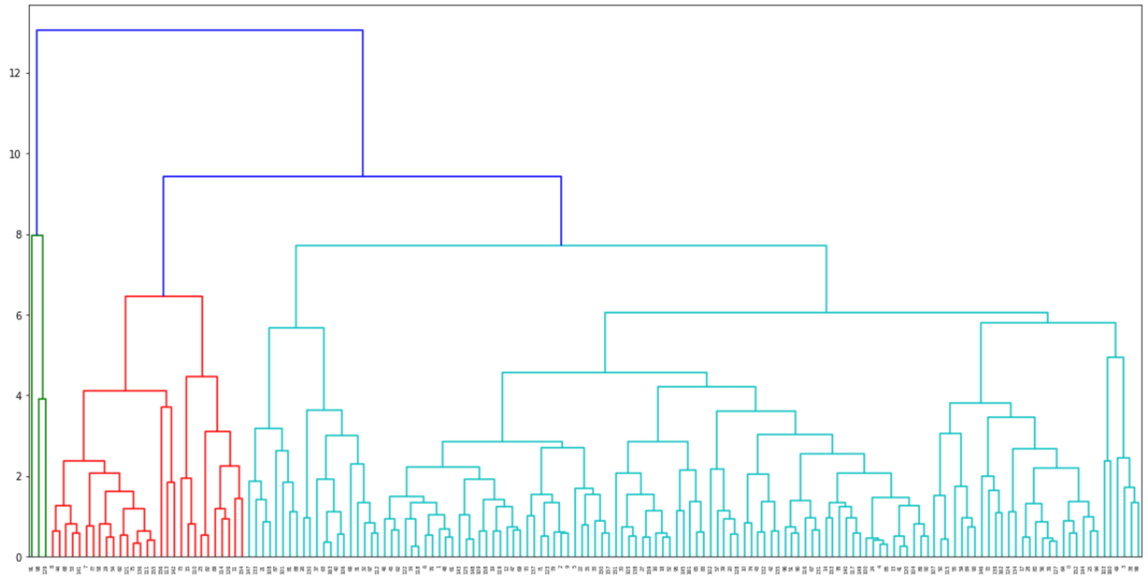
Hierarchical clustering

Complete linkage

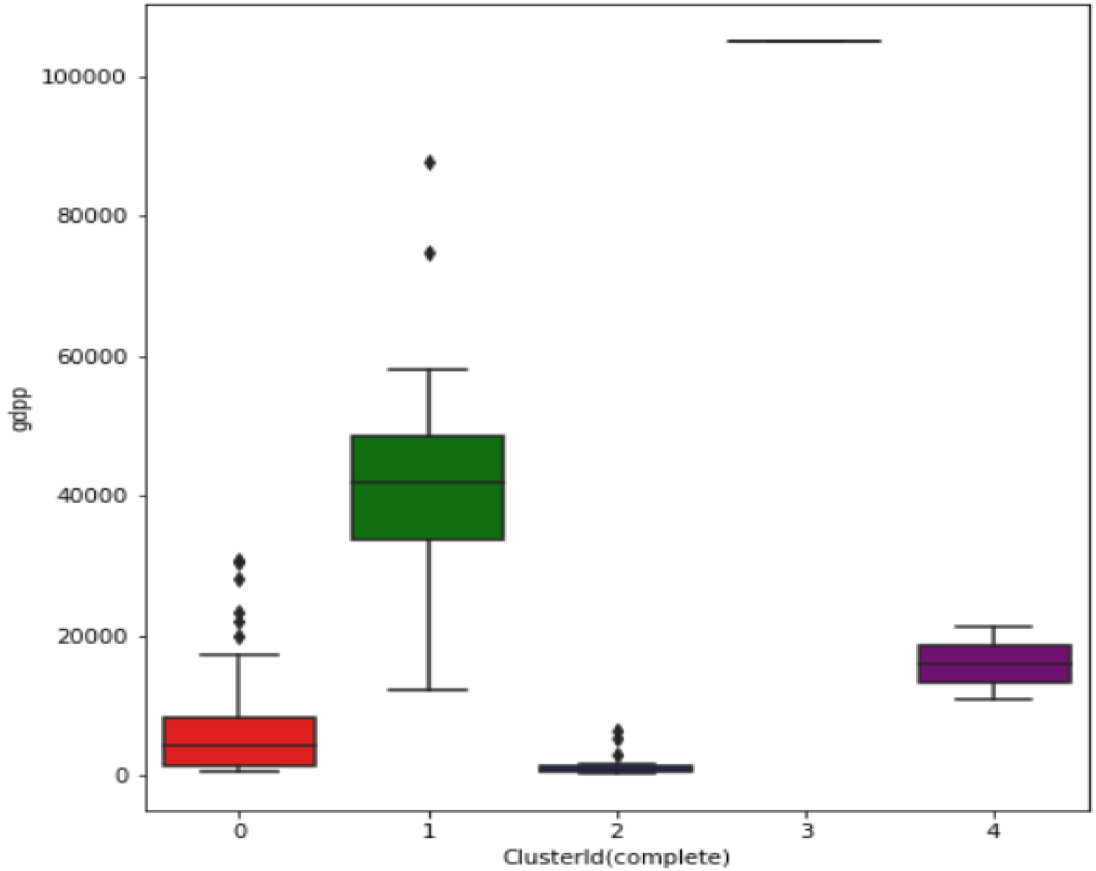
Cluster plot



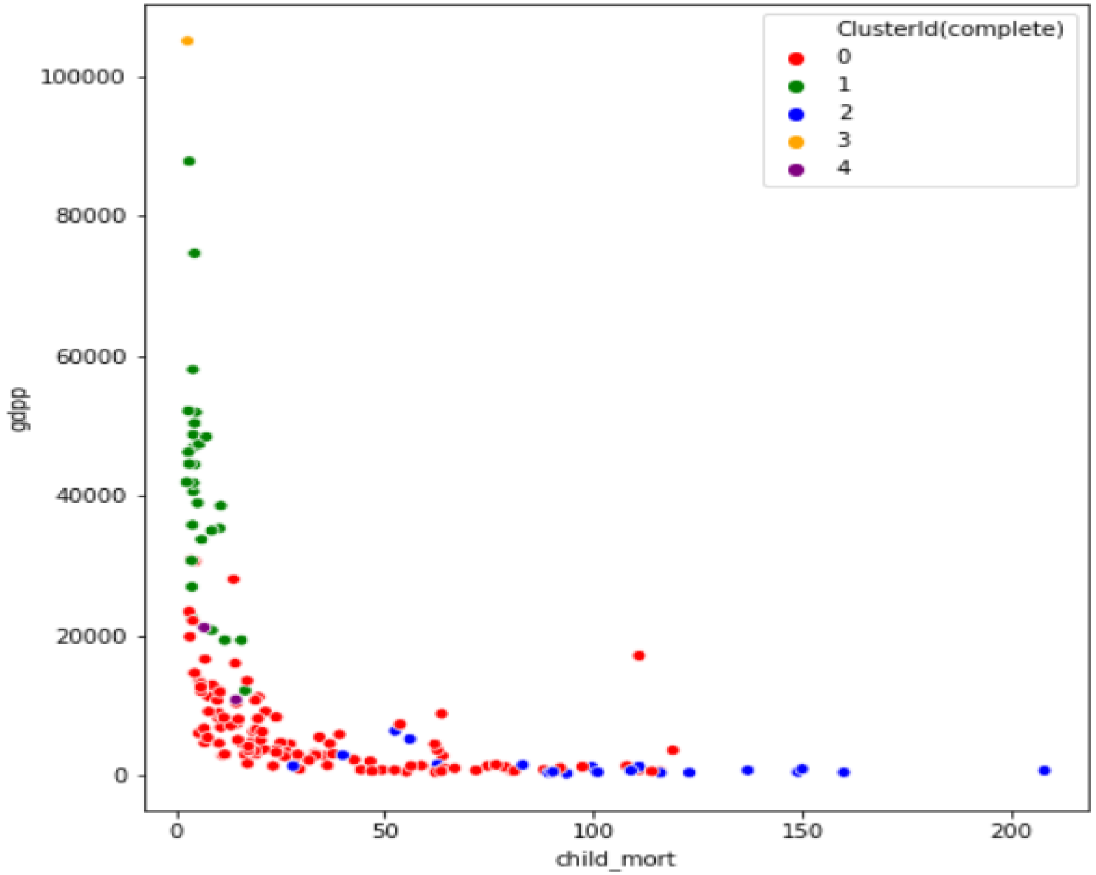
Dendrogram plot



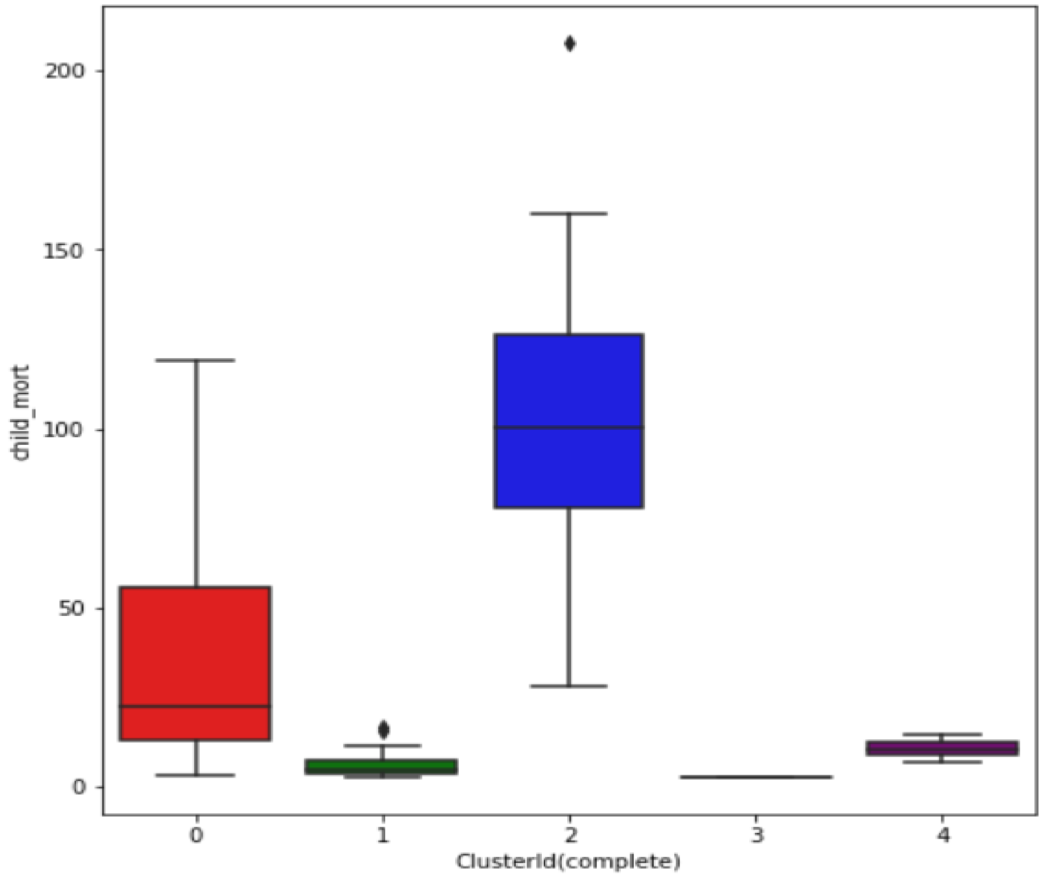
Box plot between cluster Id(complete) & GDP per capita



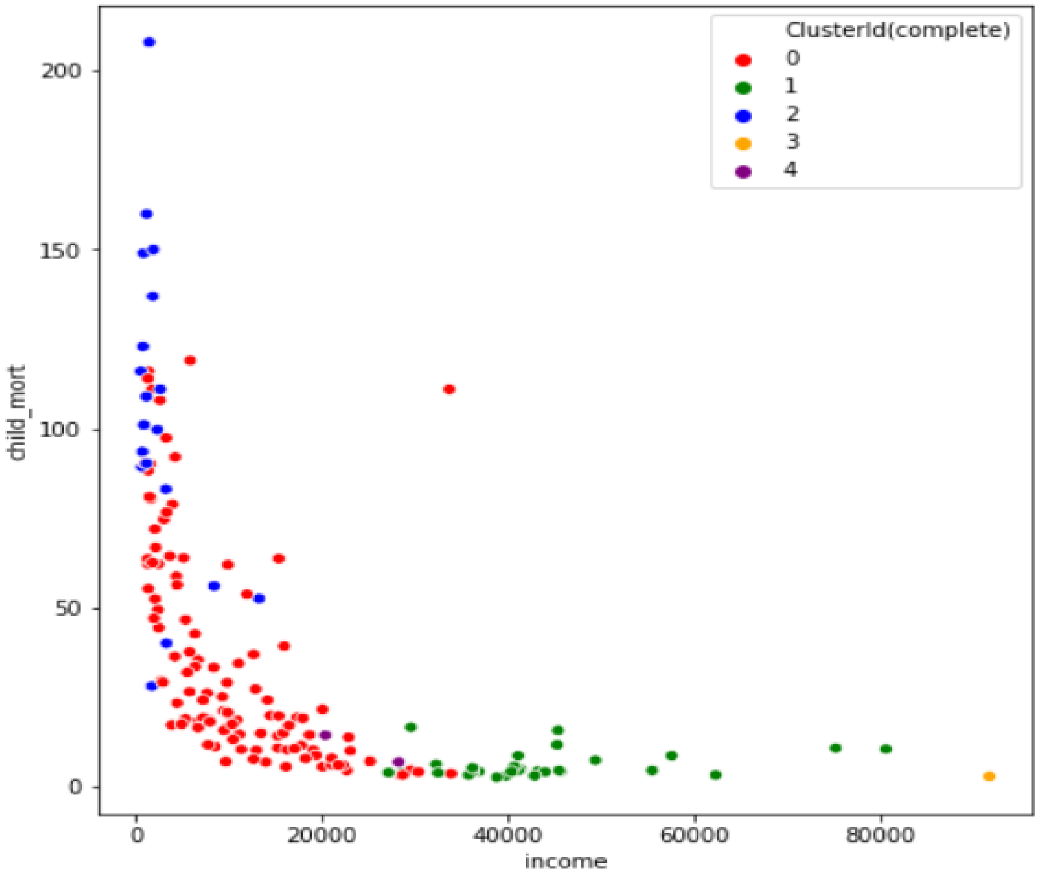
Box plot between child mortality & GDP per capita



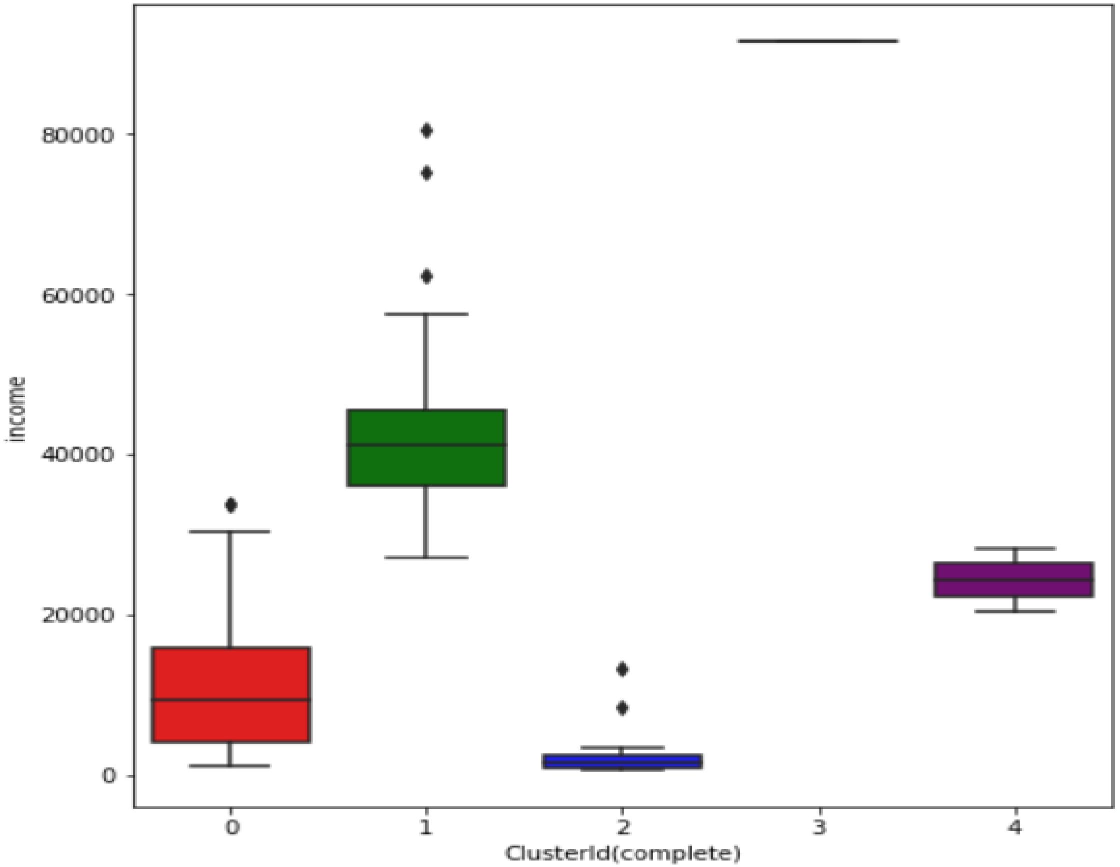
Box plot between cluster Id(complete) & GDP per capita



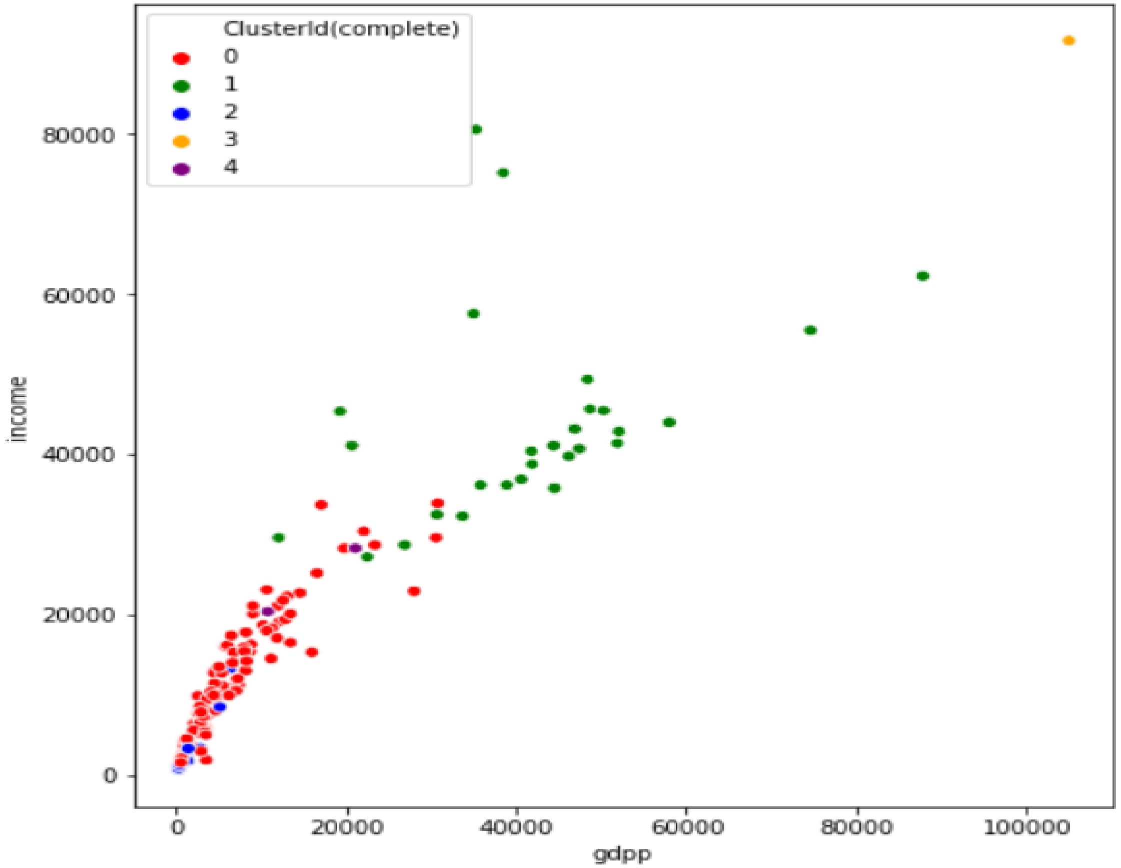
Box plot between child mortality & Income

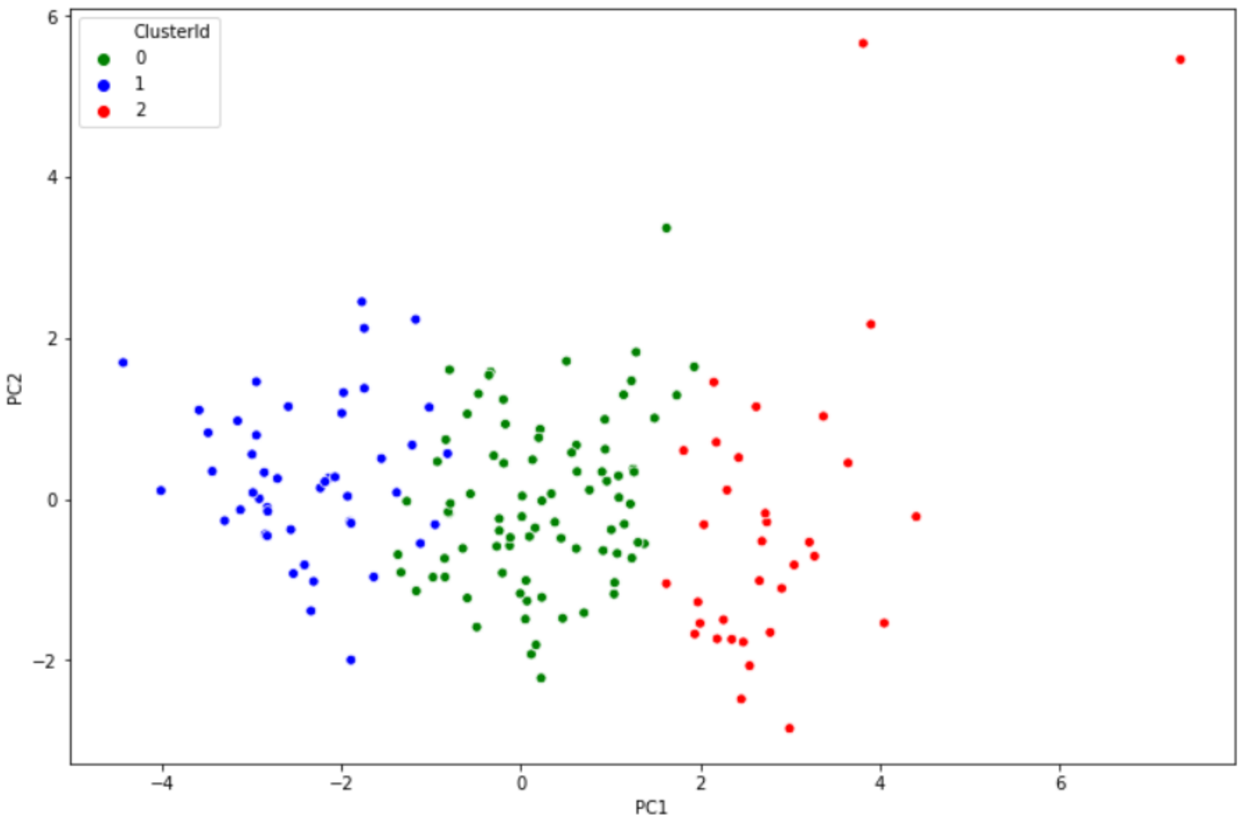


Box plot between cluster Id(complete) & income



Box plot between GDP per capita & income





Desired method

Desired result was produced using K-Means on data which is treated for outliers, scaled, transformed using PCA and then again treated for outliers.

List of countries that are in direst need for aid

serial.no	Country	serial.no	Country
1	Congo Dem. Rep.	25	Chad
2	Liberia	26	Tanzania
3	Burundi	27	Senegal
4	Niger	28	Lesotho
5	Central African Republic	29	Kenya
6	Mozambique	30	Cameroon
7	Malawi	31	Cote d'Ivoire
8	Guinea	32	Ghana
9	Togo	33	Zambia
10	Sierra Leone	34	Mauritania
11	Rwanda	35	Sudan
12	Madagascar	36	Myanmar
13	Guinea-Bissau	37	Lao
14	Comoros	38	Pakistan
15	Eritrea	39	Yemen
16	Burkina Faso	40	Congo, Rep.
17	Haiti	41	Angola
18	Uganda	42	Namibia
19	Afghanistan	43	South Africa
20	Gambia	44	Iraq
21	Kiribati	45	Botswana
22	Benin	46	Gabon
23	Timor-Leste	47	Equatorial Guinea
24	Mali		

Country are sorted based on low income and GDP per captia and high child mortality