Clustering and PCA Assignment

Author: Rudrakanta Ghosh

Applicant ID: APFE19803179

Scenario

- HELP International, an international humanitarian NGO, 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.
- HELP international have been able to raise around $ 10 million.
- CEO of the NGO needs to decide how to use this money strategically and effectively.
- We need to help the NGO in making this decision which is mostly related to choosing the countries that are in the direst need of aid

Data Set Available

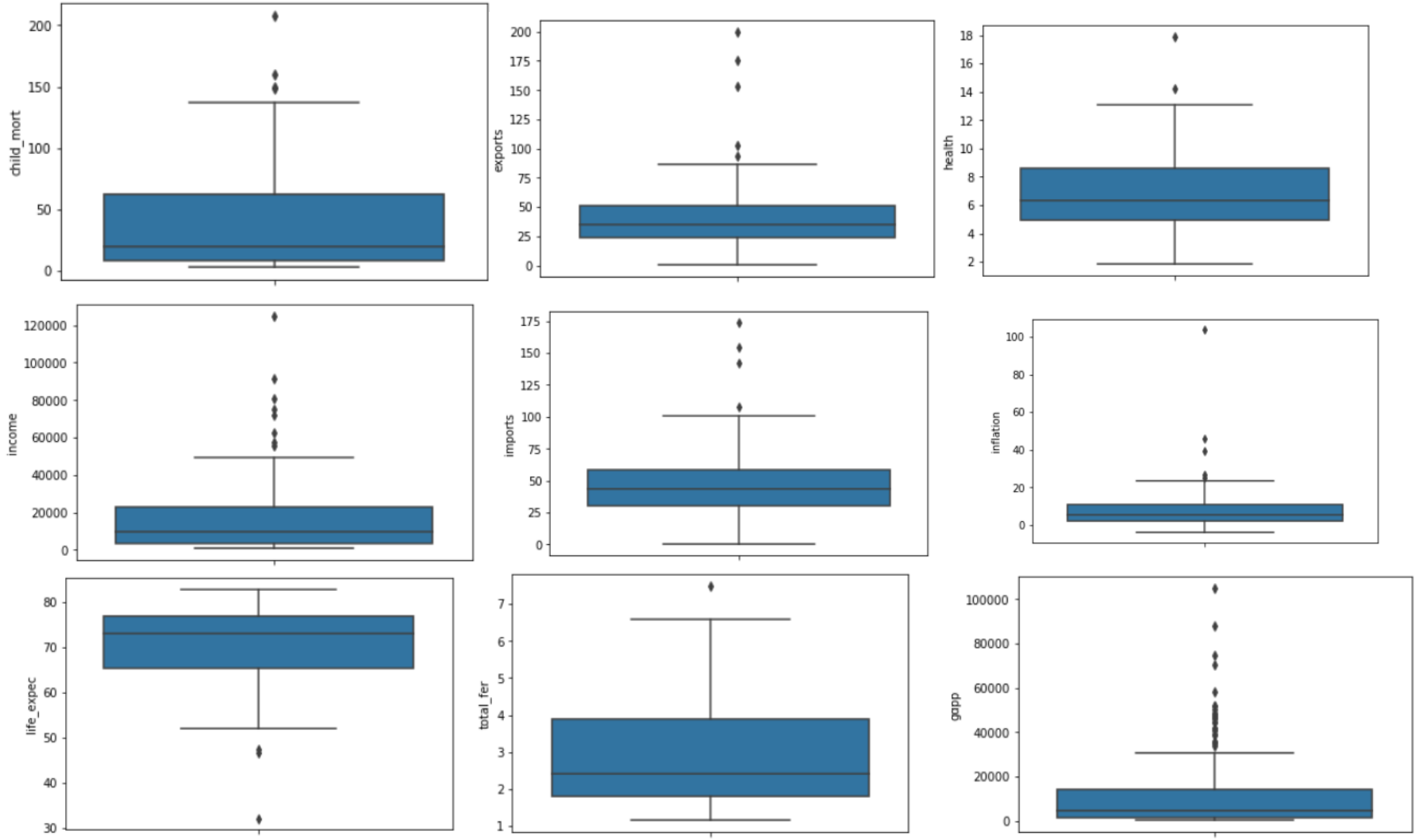
country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
Afghanistan	90.2	10.0	7.58	44.9	1610	9.44	56.2	5.82	553
Albania	16.6	28.0	6.55	48.6	9930	4.49	76.3	1.65	4090
Algeria	27.3	38.4	4.17	31.4	12900	16.10	76.5	2.89	4460
Angola	119.0	62.3	2.85	42.9	5900	22.40	60.1	6.16	3530
Antigua and Barbuda	10.3	45.5	6.03	58.9	19100	1.44	76.8	2.13	12200

	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
count	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000
mean	38.270060	41.108976	6.815689	46.890215	17144.688623	7.781832	70.555689	2.947964	12964.155689
std	40.328931	27.412010	2.746837	24.209589	19278.067698	10.570704	8.893172	1.513848	18328.704809
min	2.600000	0.109000	1.810000	0.065900	609.000000	-4.210000	32.100000	1.150000	231.000000
25%	8.250000	23.800000	4.920000	30.200000	3355.000000	1.810000	65.300000	1.795000	1330.000000
50%	19.300000	35.000000	6.320000	43.300000	9960.000000	5.390000	73.100000	2.410000	4660.000000
75%	62.100000	51.350000	8.600000	58.750000	22800.000000	10.750000	76.800000	3.880000	14050.000000
max	208.000000	200.000000	17.900000	174.000000	125000.000000	104.000000	82.800000	7.490000	105000.000000

Observation on Data Set

- The data set doesn't have null values in any of the columns
- Imputing data for missing values is not required
- Data type of all the columns(except feature variable country) are numeric(int64 and float64). Data type modification is not required for doing PCA and clustering

Outlier Analysis



Outlier Analysis

Observation:

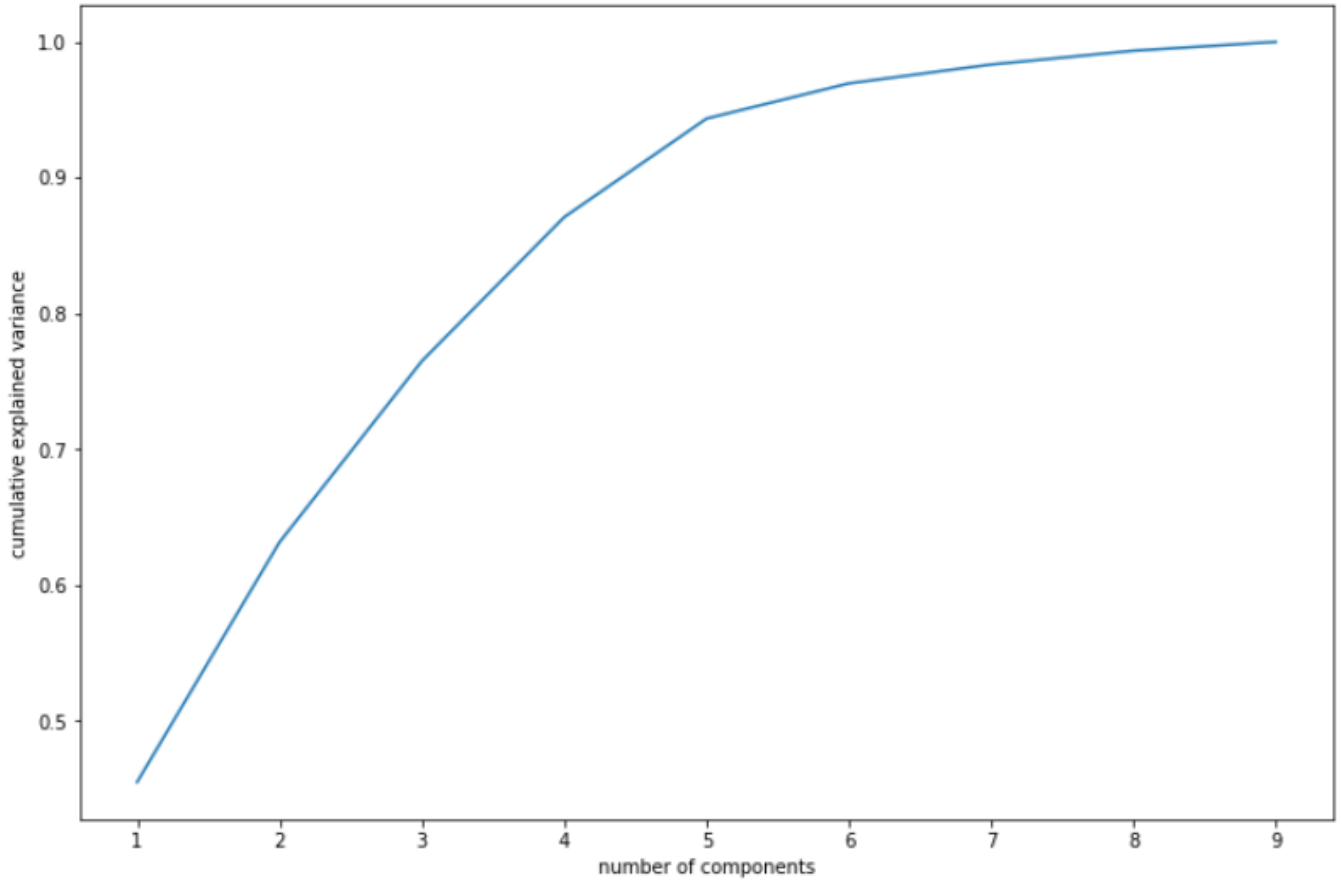
- Statistically all the columns are having outliers
- Columns like child mortality, imports, inflation are having significant number of outliers on top. Also there are outliers on bottom on life expectency. These signifies that there are few countries which are far backward economically and socially
- Columns like exports, health, income, gdpp are having significant number of outliers on top signifying there are few countries which are far ahead economically and socially compared to others

Actions on Outliers

- Since the objective of the NGO is to allocate money to countries who are dire need, we will not remove the outliers who are lagging behind economically and socially
- We will remove the outliers who are much ahead econimically and socially. Otherwise these countries will push away other countries from their cluster and we might get high number of countries in the cluster we are looking for

Principal Component Analysis

Scree Plot with Variance of Principal Components

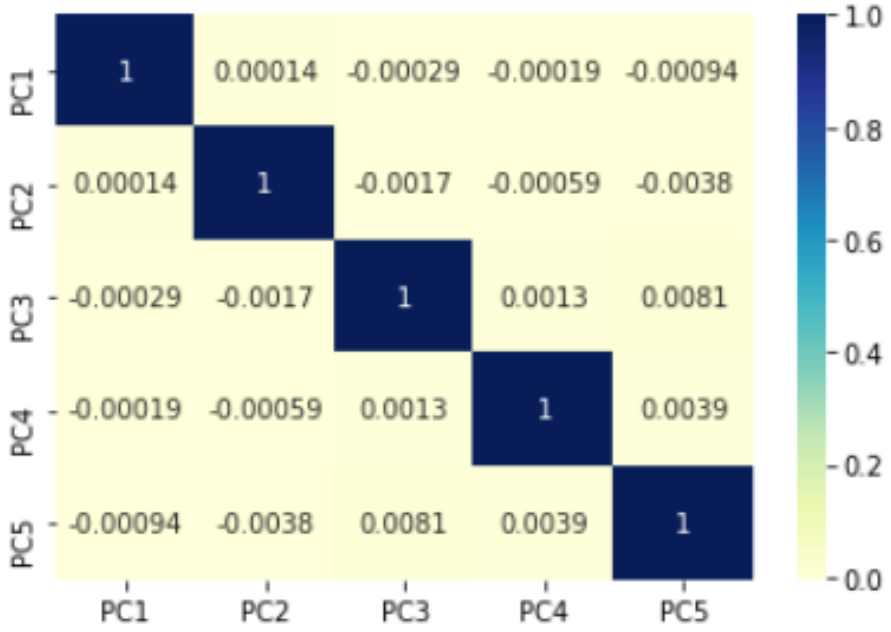


Principal Component Analysis

Observation:

- Scree plot suggests that the first 5 principal components alone can explain around 95% of the data
- We can continue with 5 principal components for doing rest of the analysis

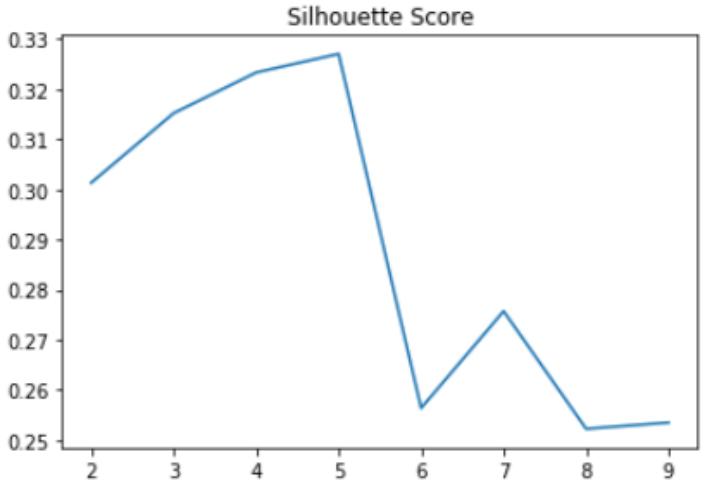
Correlation coefficients among PC’s



Observation: As expected, there is no linear dependency among the principal components

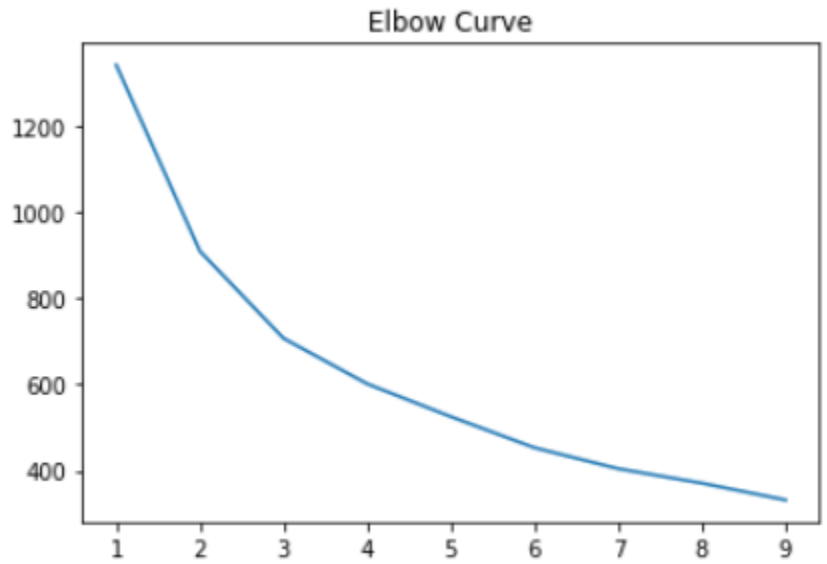
K-Means Clustering

Hopkins Statistic Score: .83 → Good score for K-Means clustering



Silhoutte score suggests that 5 can be an Ideal number of cluster for K-Means clustering method.
Lets check Elbow curve if we can have a less number of cluster comparing the data volume

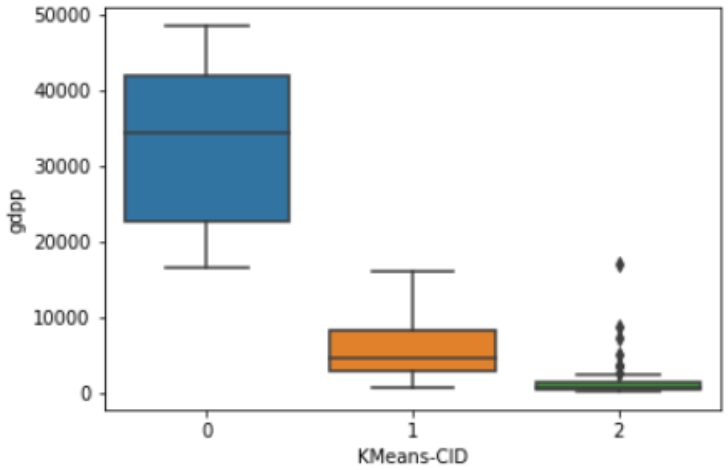
K-Means Clustering

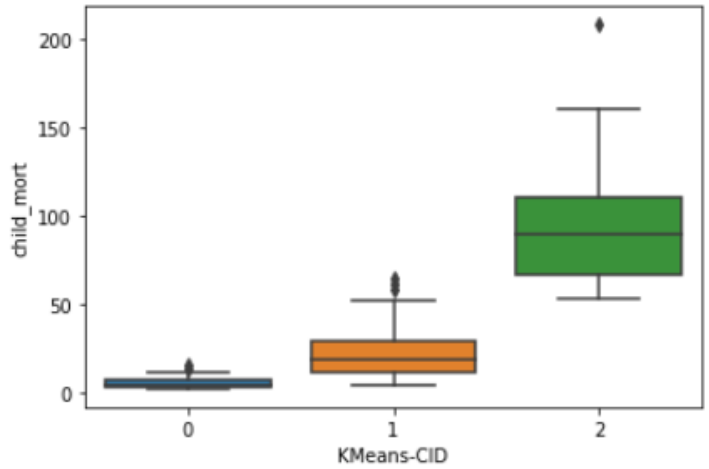


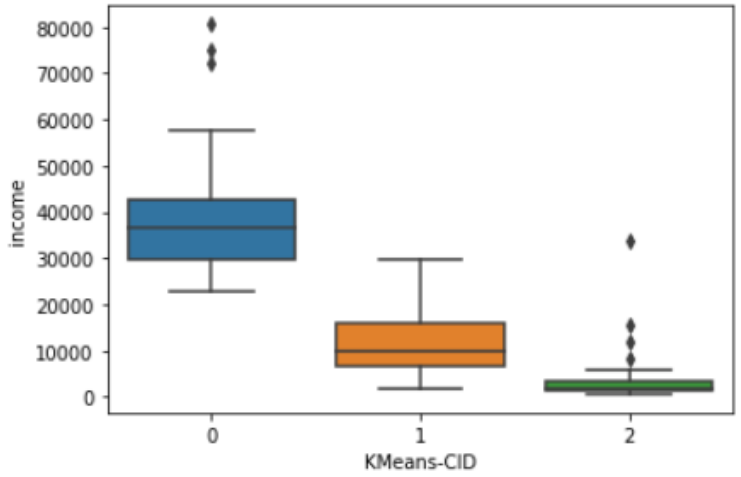
Elbow curve suggests that 3 can be an Ideal number of cluster for K-Means clustering method.

We will go with 3 cluster for K-Means clustering technique.

K-Means: Analysis of GDP, Child Mortality and Income for each cluster







K-Means Clustering

Observation

- Cluster 2 countries are of low GDP, high child mortality and low income group
- Cluster 0 countries are of high GDP, low child mortality and high income group
- Cluster 1 countries are of medium GDP, child mortality and income

K-Means Clustering Conclusion:

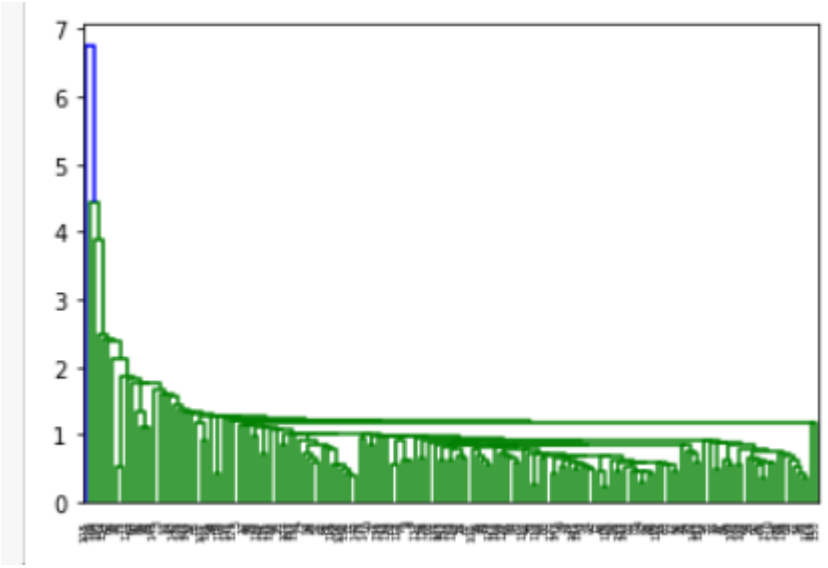
- Cluster 2 countries are in dire need of financial need. The NGO should give the aid to them.

List of countries needing aid - as per K-Means clustering

Afghanistan	Gabon	Mauritania
Angola	Gambia	Mozambique
Benin	Ghana	Namibia
BurkinaFaso	Guinea	Niger
Burundi	Guinea-Bissau	Nigeria
Cameroon	Haiti	Pakistan
CentralAfricanRepublic	Kenya	Rwanda
Chad	Kiribati	Senegal
Comoros	Lao	SierraLeone
Congo,Dem.Rep.	Lesotho	SouthAfrica
Congo,Rep.	Liberia	Sudan
Coted'Ivoire	Madagascar	Tanzania
EquatorialGuinea	Malawi	Timor-Leste
Eritrea	Mali	Togo
Zambia	Uganda	Yemen

Hierarchical Clustering

Single Linkage

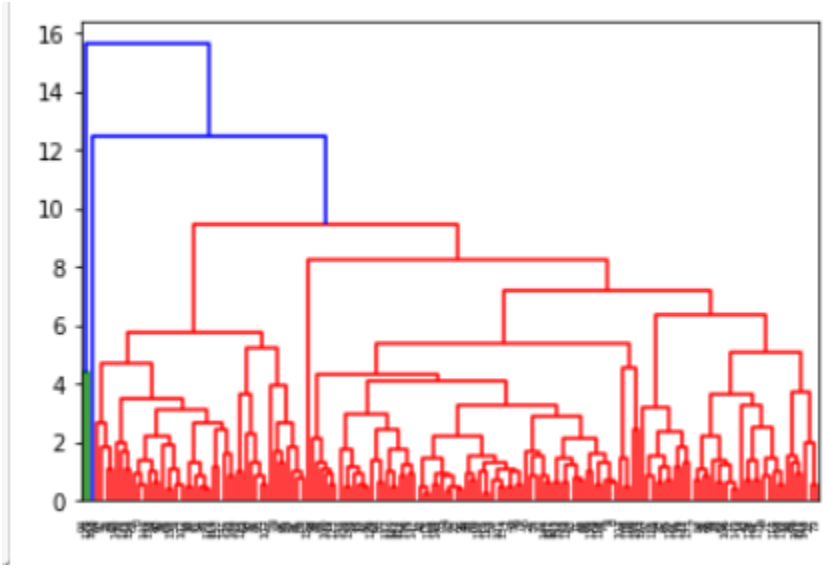


Observation

- The plot is not clear enough to cut the tree and decide the number of cluster
- Single Linkage is not good enough. We should be doing using complete linkage

Hierarchical Clustering

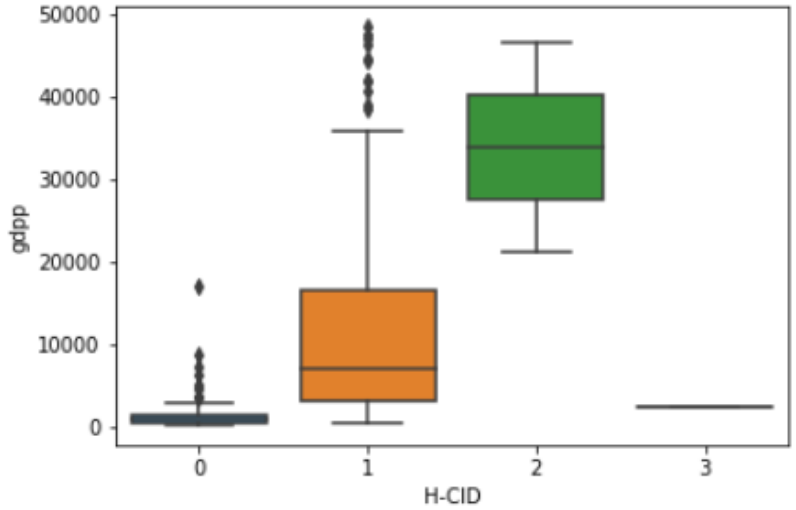
Complete Linkage

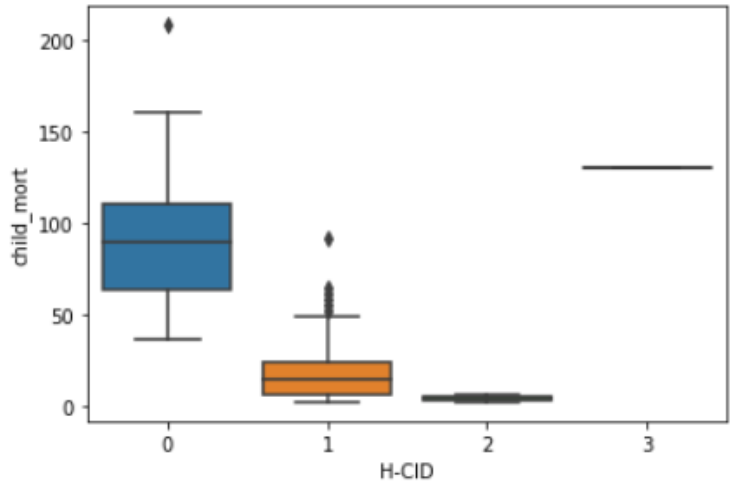


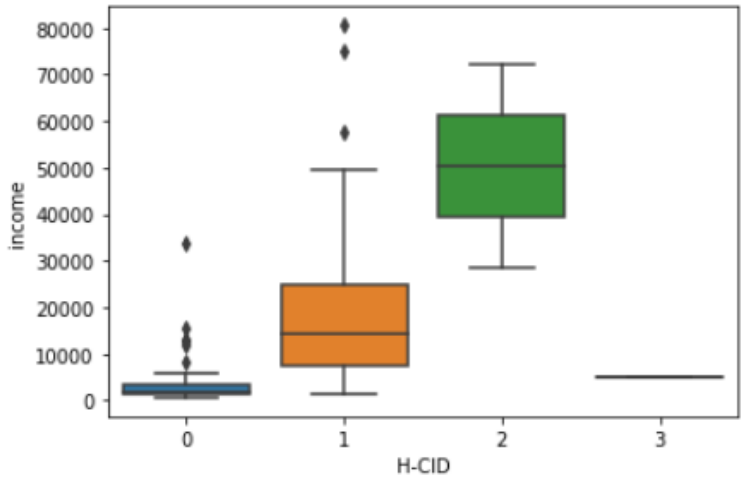
Observation:

- Complete Linkage shows a possibility of cutting the tree at 9 resulting in 3 cluster
- With 3 cluster, distribute the data in a skewed form with most of the countries going in one cluster
- Complete Linake also shows a possibility of cutting the tree between 7 and 8, resulting in 4 cluster
- With 4 cluster, skewness of distribution seems to be overcomed

Hierarchical Clustering: Analysis of GDP, child mortality and income for each cluster







Hierarchical Clustering

Observation

- Cluster 0 and 3 countries are of low GDP, high child mortality and low income group
- Cluster 2 countries are of high GDP, low child mortality and high income group
- Cluster 1 countries are of medium GDP, child mortality and income

Hierarchical Clustering Conclusion:

- Cluster 0 and 3 countries are in dire need of financial need. The NGO should give the aid to them.

List of countries needing aid - as per Hierarchical clustering

Afghanistan
Angola
Benin
Botswana
BurkinaFaso
Burundi
Cameroon
CentralAfricanRepublic
Chad
Comoros
Congo,Dem.Rep.
Congo,Rep.
Coted'Ivoire
EquatorialGuinea
Gabon
Gambia
Ghana
Guinea
Guinea-Bissau
Haiti
Iraq
Kenya
Kiribati
Lao
Lesotho
Liberia
Madagascar
Malawi
Mali
Mauritania
Micronesia,Fed.Sts.
Mozambique
Namibia
Niger
Nigeria
Rwanda
Senegal
SierraLeone
SouthAfrica
Sudan
Tanzania
Timor-Leste
Togo
Uganda
Yemen
Zambia