Assignment-4
Clustering and PCA based analysis on national level socio-economic factors

Presented by

Prashant Chadha

Problem statement

The goal is to analyze the socio-economic parameters given for each country in the dataset and identify list of countries that require aid the most. Analysis to be done using clustering (K-means and hierarchical) as well as principal component analysis or PCA.

Solution

Following are the steps that were carried out during analysis:

- Perform EDA on the given dataset to check for data inconsistencies such as missing values, duplicate entries, formatting issues, outlier analysis, etc.
- Scale the numeric columns via StandardScaler so that all variables are within range of each other. This helps in accurate PCA and clustering processes
- Plot correlation heat map to analyze the correlation between the numeric variables
- Perform PCA to derive the primary components that explain 95% of variance
- Run both K-means and hierarchical clustering algorithms on the principal components to form clusters
- Select the cluster that most likely contains countries that fit the required business case and then shortlist few countries that require aid the most

Dataset snapshot

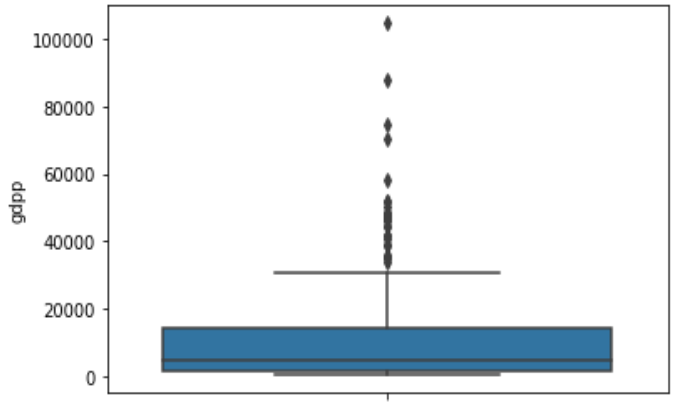
	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
0	Afghanistan	90.2	10.0	7.58	44.9	1610	9.44	56.2	5.82	553
1	Albania	16.6	28.0	6.55	48.6	9930	4.49	76.3	1.65	4090
2	Algeria	27.3	38.4	4.17	31.4	12900	16.10	76.5	2.89	4460
3	Angola	119.0	62.3	2.85	42.9	5900	22.40	60.1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

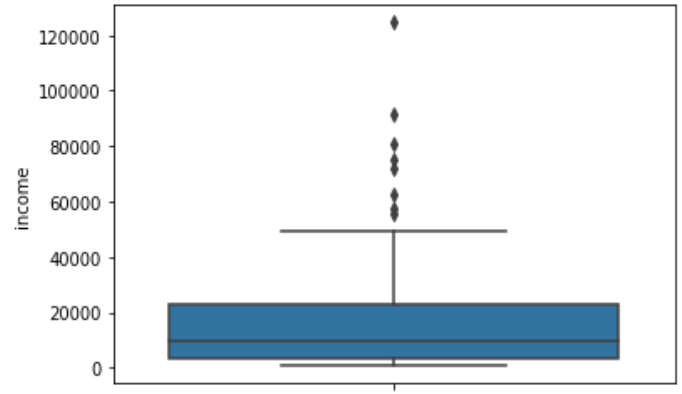
The given dataset, as can be seen from above, consists the following columns:

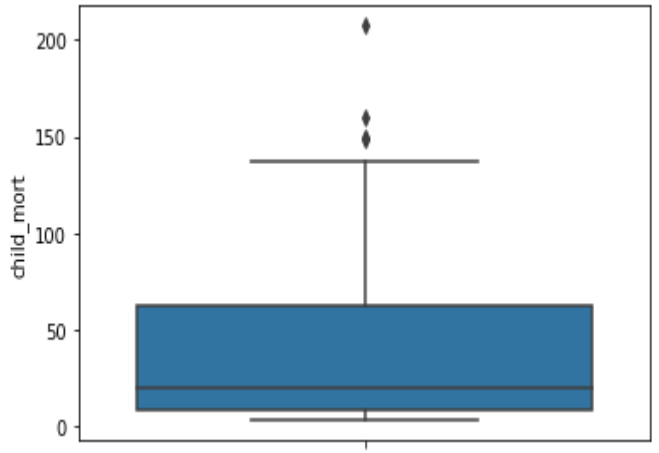
- country: This column consists of countries for which the socio-economic parameters are given. There are total of 167 countries
- child_mort: Child mortality given as death of children under 5 years of age per 1000 live births
- exports: Exports of goods and services. Given as %age of the Total GDP
- imports: Imports of goods and services. Given as %age of the Total GDP
- health: Total health spending as %age of Total GDP
- income: Net income per person
- inflation: The measurement of the annual growth rate of the Total GDP
- life_ expec: The average number of years a new born child would live if the current mortality patterns are to remain the same
- total_fer: The number of children that would be born to each woman if the current age-fertility rates remain the same
- gdpp: The GDP per capita. Calculated as the Total GDP divided by the total population

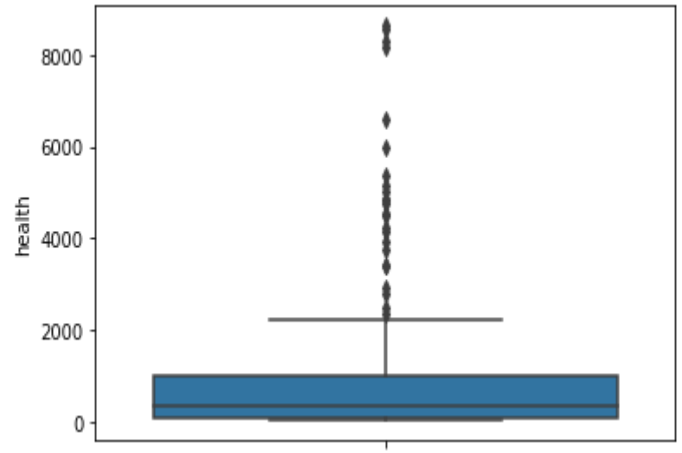
EDA and data preparation Report

- There were no missing values found in the dataset
- There are no duplicate entries in the dataset as well
- Converted the 'exports', 'health', 'imports' from percentage values, relative to total GDP, to actual values. I had to consider GDP per capita only instead of total GDP as that's the only detail available.
- Outlier analysis: It can be seen that outliers do exist in the variables for which distribution has been visualized using box plot. Outliers will be removed after scaling and PCA analysis
- Numeric variables were scaled using StandardScaler to normalise the data



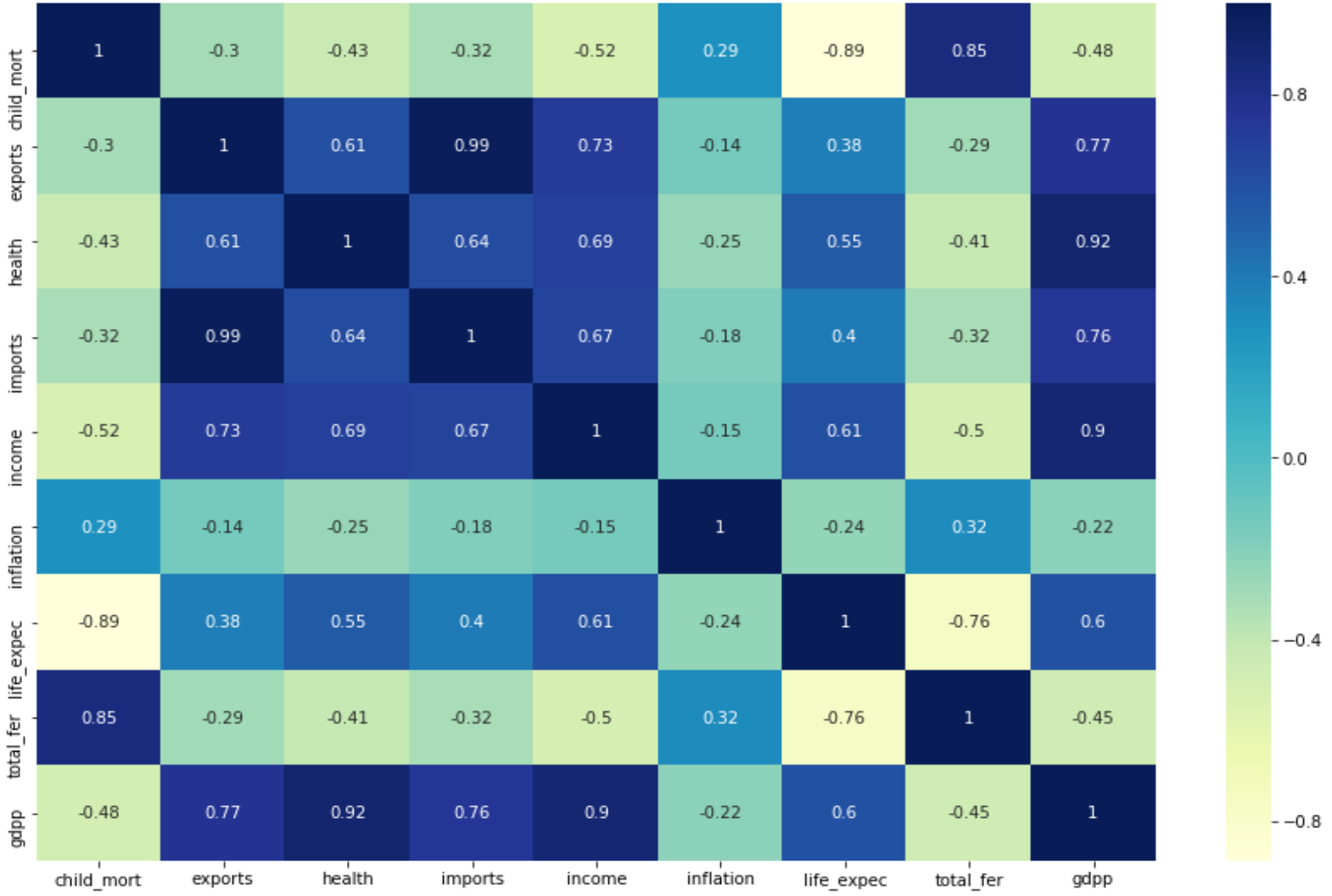




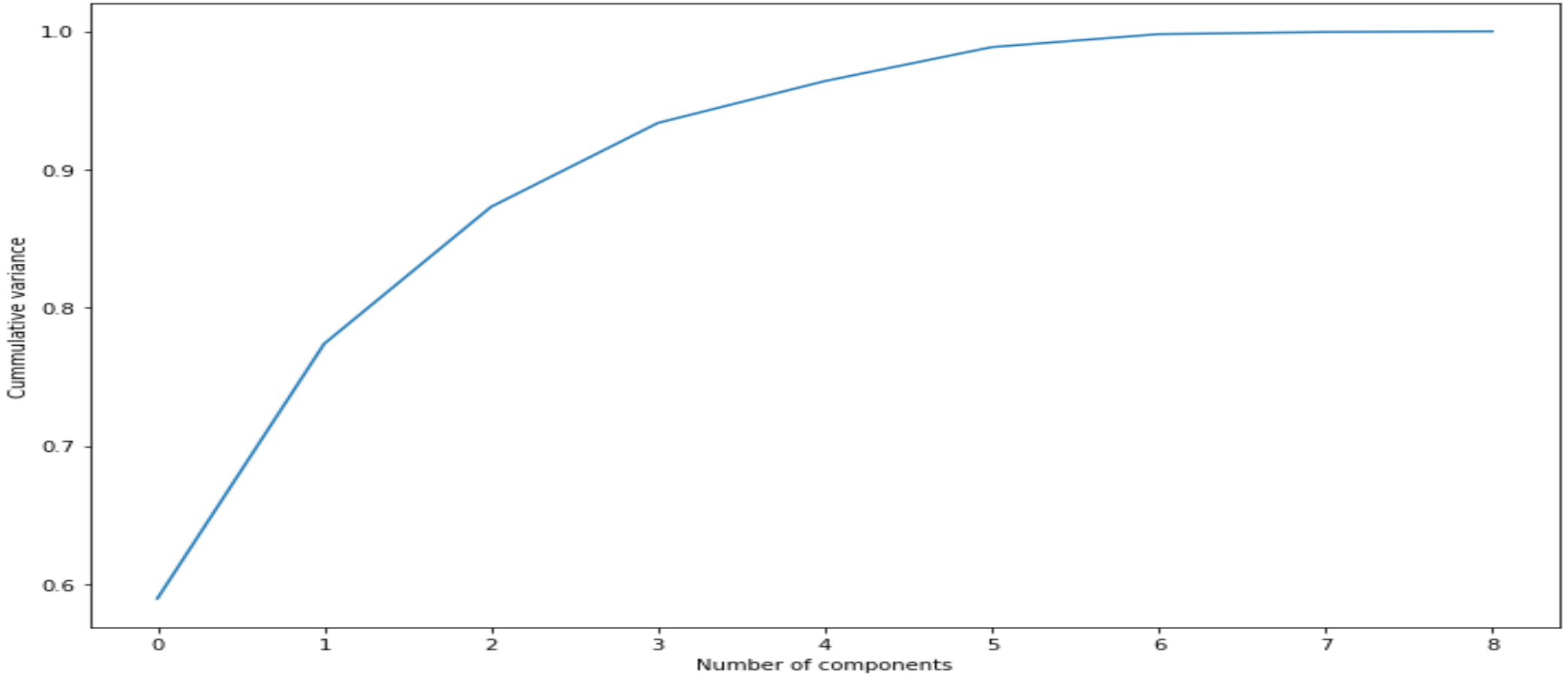


Correlation Heat map

- While it was expected that 'exports', 'health' and 'imports' would be highly correlated with 'gdpp', 'income' and 'life_expec' are the other two major correlated variables with 'gdpp'.
- Therefore, 'gdpp' and 'income' can together explain enough variance for other correlated parameters as well.
- Also, 'life_expec' is highly negatively correlated to 'child_mort' and 'total_fer', which goes with the definition of these columns as well.



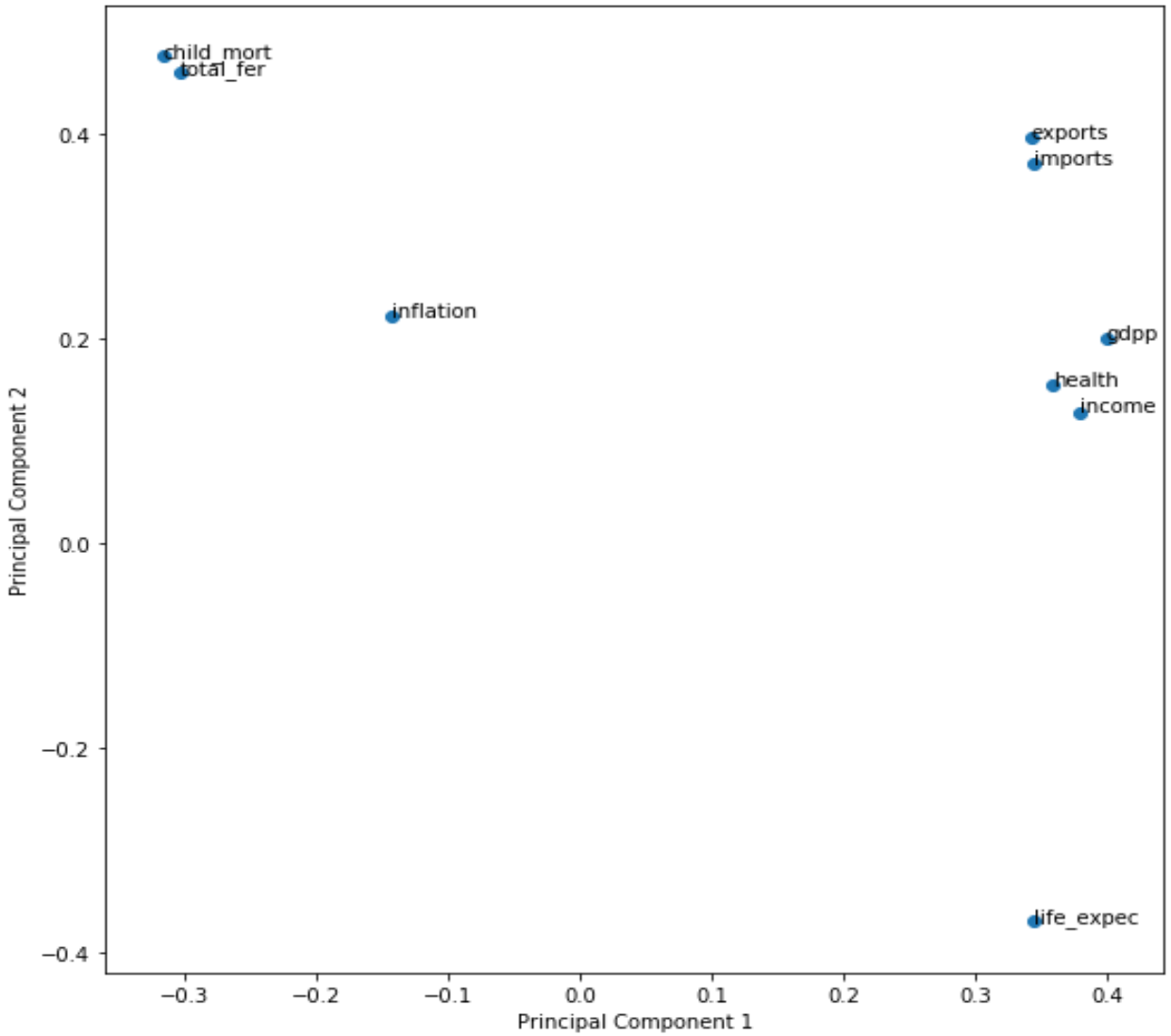
PCA Report – 1: Number of Principal components selection



- As can be seen from above graph, five principal components (PCs) combined explain variance of approximately 95%
- Therefore, I considered the PCs to be five and ran the clustering algorithms on these PCs

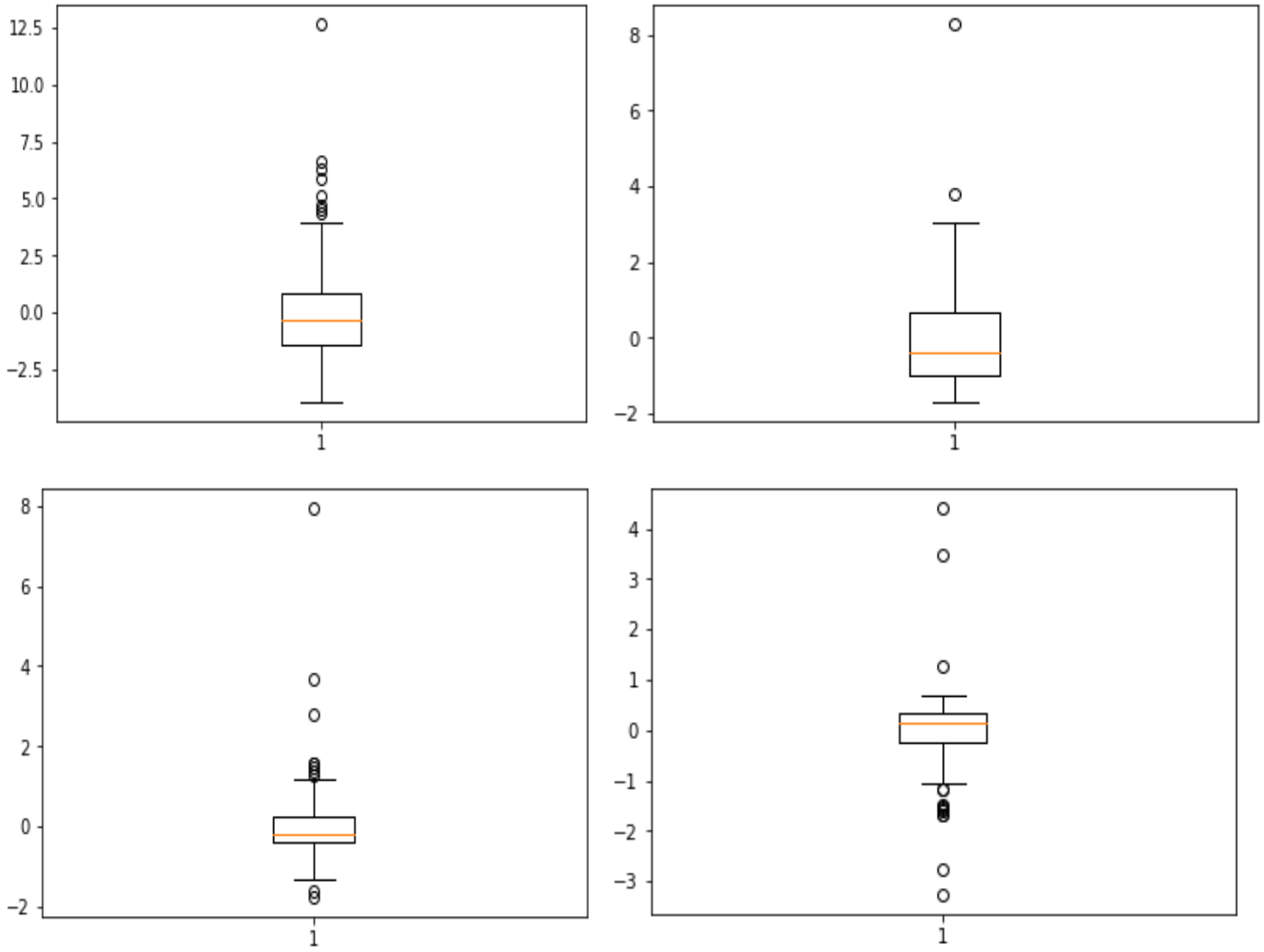
PCA Report – 2: Scatterplot between the first two PCs, with points representing the original variables

- It can be seen from the plot that variables such as ‘child_mort’ and ‘total_fer’ are in the direction of principal component-2 (PC2).
- The variables such as ‘gdpp’, ‘health’, ‘income’ and ‘life_expec’ are in the PC1.
- Variables ‘imports’ and ‘exports’ have similar PC1 and PC2 values.
- From similar scatterplots between rest of the first 5 PCs combinations, it can be verified that most of the variance is contributed by these primary variables .



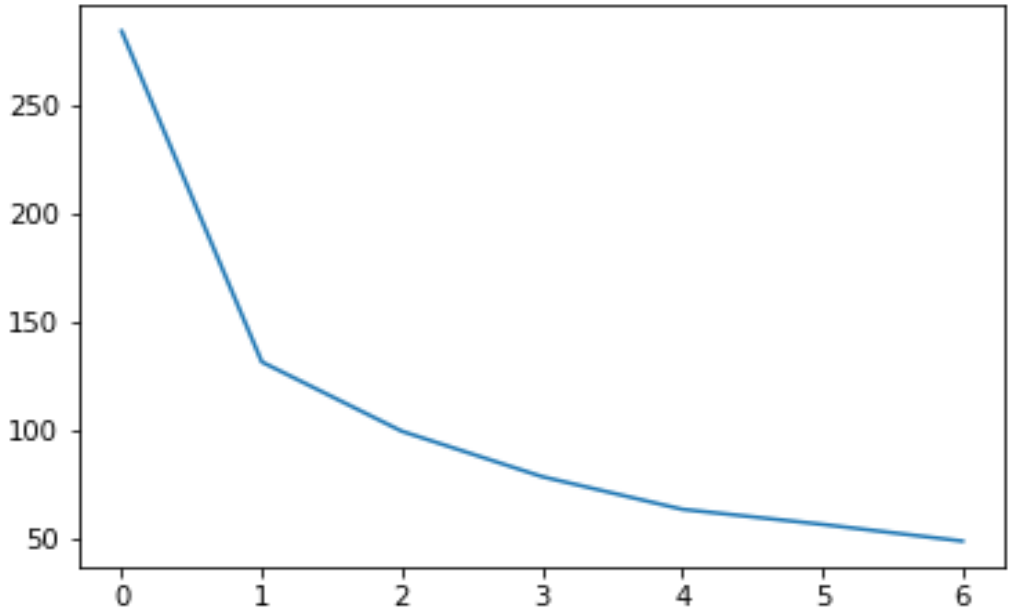
PCA Report – 3: Outlier removal

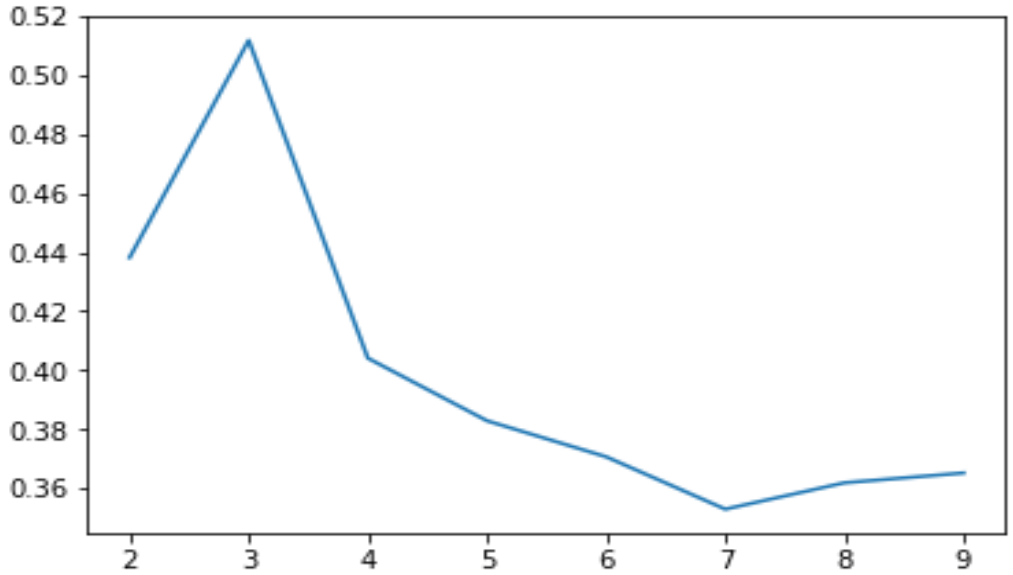
- The boxplots for the first four PCs are shown on the right and outliers can be clearly seen. PC5, not shown here, too had outliers in its boxplot.
- Though we remove the outliers above 95 percentile and below 5 quantile , I removed the outliers only above 95 percentile for all components except, 'PC2',
- As we are interested in cluster of countries with lower than usual values, as these would probably consist of countries that require the funding most.
- 'PC2' explains the highest variance in the direction of 'child_mort' which would be high for countries that are in need of aids. Therefore, will remove the lower than usual quantile values for 'PC2'.



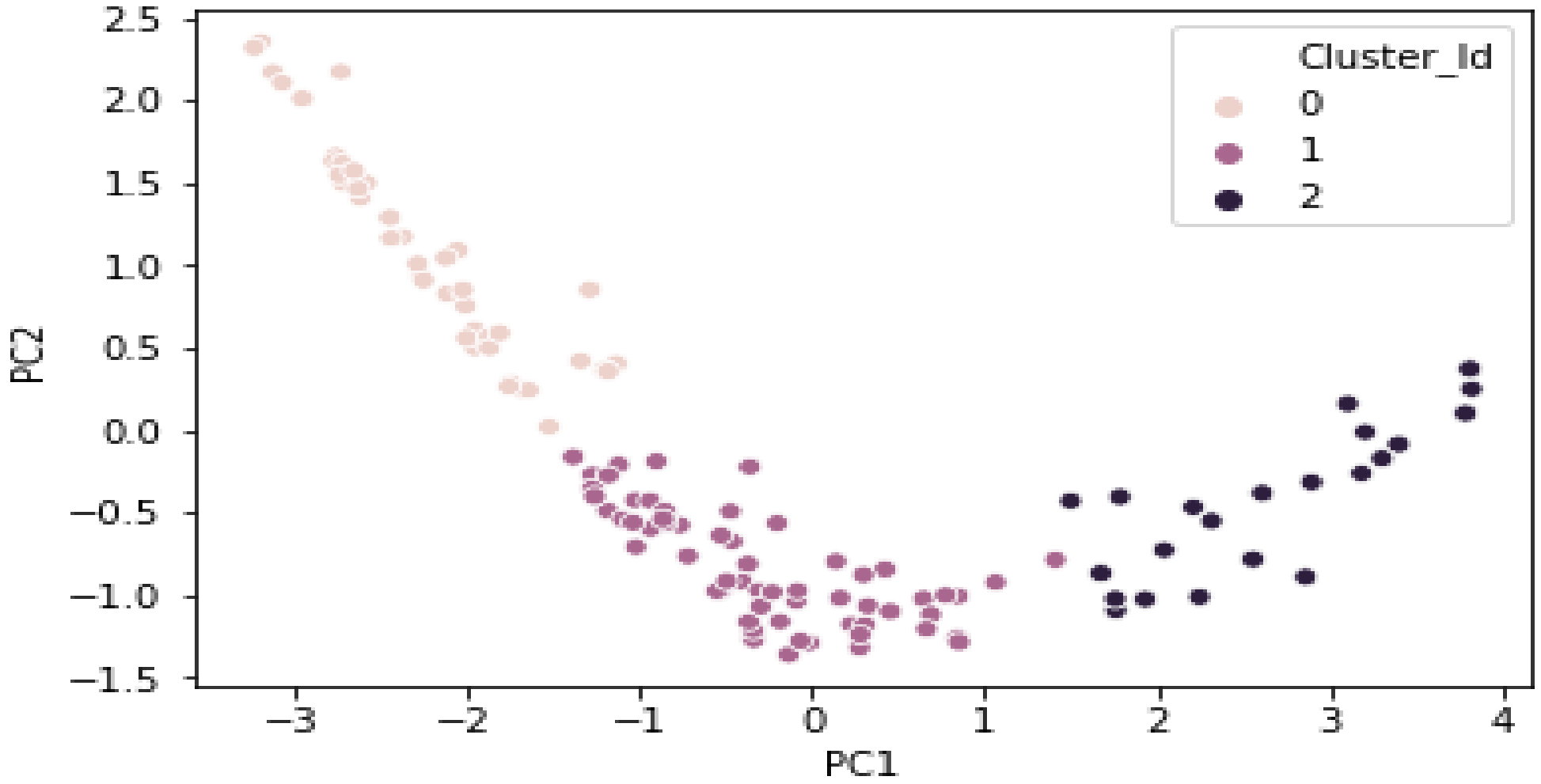
K-means clustering-1

- First step was to calculate ‘Hopkins’ score to verify whether the dataframe consisting of five PCs (linear transformations of respective original variables)
- The score was calculated to be 0.84 or 84% which is well above accepted baseline score for cluster analysis acceptability of data
- Next, elbow curve was plotted as shown by the graph in the top right corner.
- Silhouette score was also plotted as shown by the graph in the bottom right corner
- Both the methods suggested number of clusters to be ‘3’ as ideal number of clusters for this case
- Cluster Ids were calculated for the three clusters using K-means algorithm
- Next, the three clusters were formed by associating respecting Cluster Ids, first to the five PCs dataframe and then to countries by merging this dataframe with parent dataframe consisting of countries and original variables



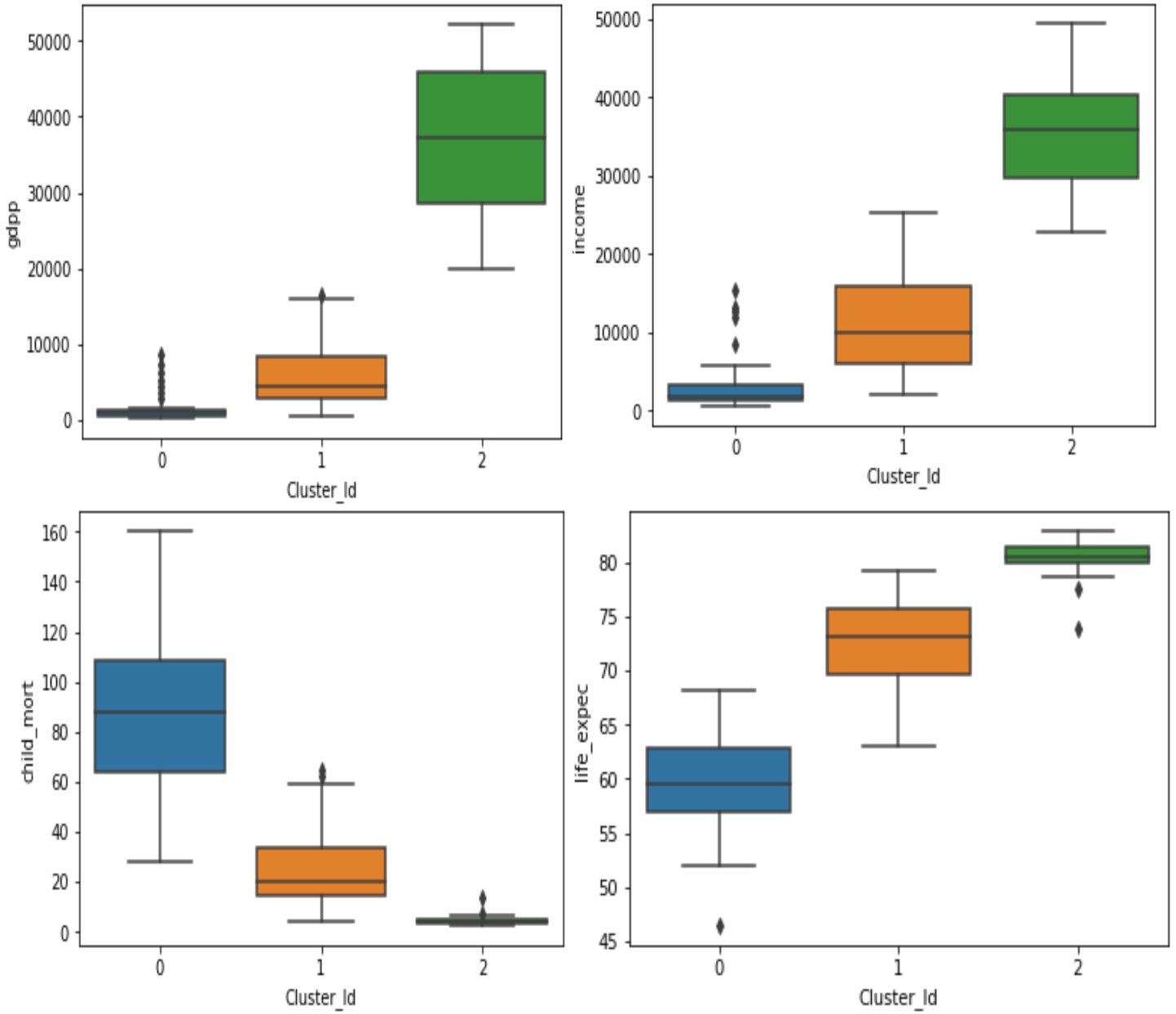


Scatterplot between PC1 and PC2, with points representing their respective clusters



K-means clustering-2: Boxplot analysis

- It can be seen from box plots that cluster -0 has the countries of interest due to the following analysis results:
- The 'gdpp' is lowest for this cluster. Therefore, the spending on 'health', 'imports' and 'exports' is least and in fact GDP per capita, given by 'gdpp' is well below acceptable limits.
- The 'child_mort' rate is highest for this cluster and again, alarmingly high, so countries within this cluster need the most attention
- The 'income' is lowest for this cluster so the spending power of an individual from countries within this cluster is lowest which means that access to basic needs is toughest for people of these countries
- Therefore, created a dataframe with countries from cluster-0 only for further analysis



K-means clustering-3: Shortlisting dire state countries

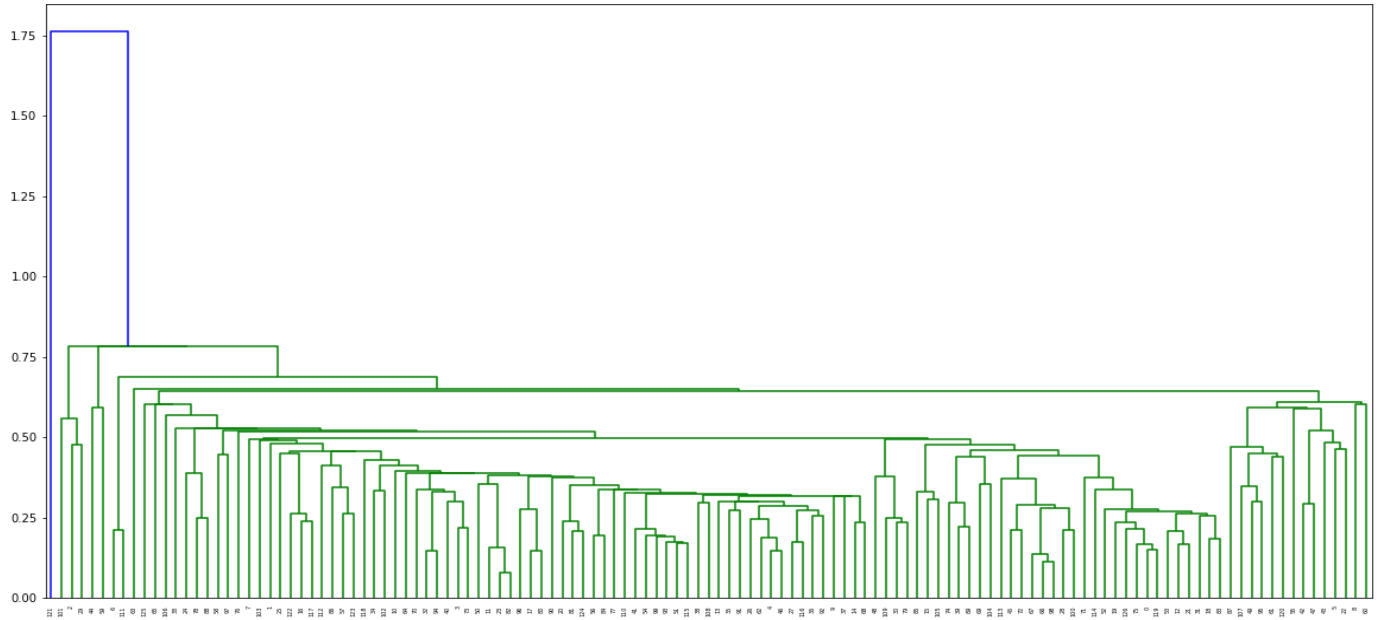
- First shortlisted the countries within the dataframe that have values of 'gdpp', 'income' lower than or equal to their mean , and, have 'child_mort' higher than or equal to the mean value.
- This resulted in reduced dataset but still had a number of countries which did not require immediate financial aid.
- Next, filtered this dataset based on 'total_fer' being greater than or equal to the mean and 'income' being lower than or equal to the mean of this dataset. This is because such countries would have the most number of children with parental income being lowest and therefore, would require financial aid most urgently.
- The resulting dataset consisted reduced number of countries however, further filtering was still possible
- Filtered the countries further based on 'child_mort’ and derived the dataframe consisting of countries that are in dire need of funds.

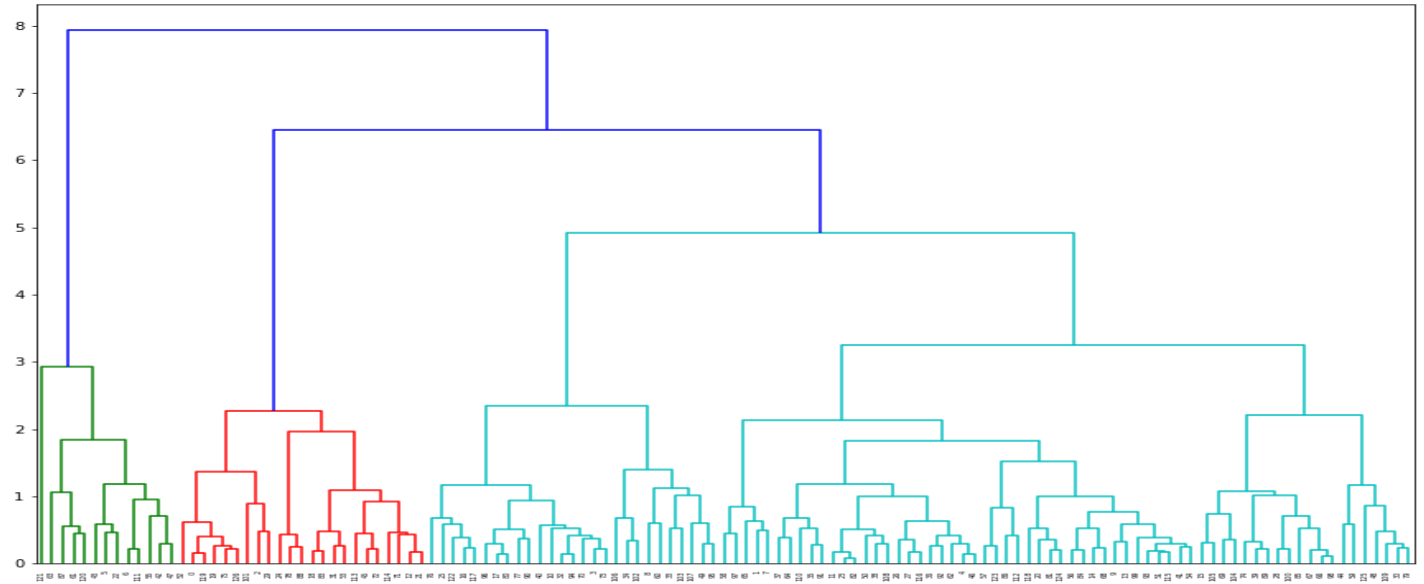
K-means clustering- 4: Final dataframe with shortlisted countries and respective parameters

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	Cluster_Id
18	Burkina Faso	116.0	110.40	38.76	170.20	1430	6.81	57.9	5.87	575	0
29	Congo, Dem. Rep.	116.0	137.27	26.42	165.66	609	20.80	57.5	6.54	334	0
52	Guinea	109.0	196.34	31.95	279.94	1190	16.10	58.0	5.34	648	0
53	Guinea-Bissau	114.0	81.50	46.50	192.54	1390	2.97	55.6	5.05	547	0
88	Niger	123.0	77.26	17.96	170.87	814	2.55	58.8	7.49	348	0
101	Sierra Leone	160.0	67.03	52.27	137.66	1220	17.20	55.0	5.20	399	0

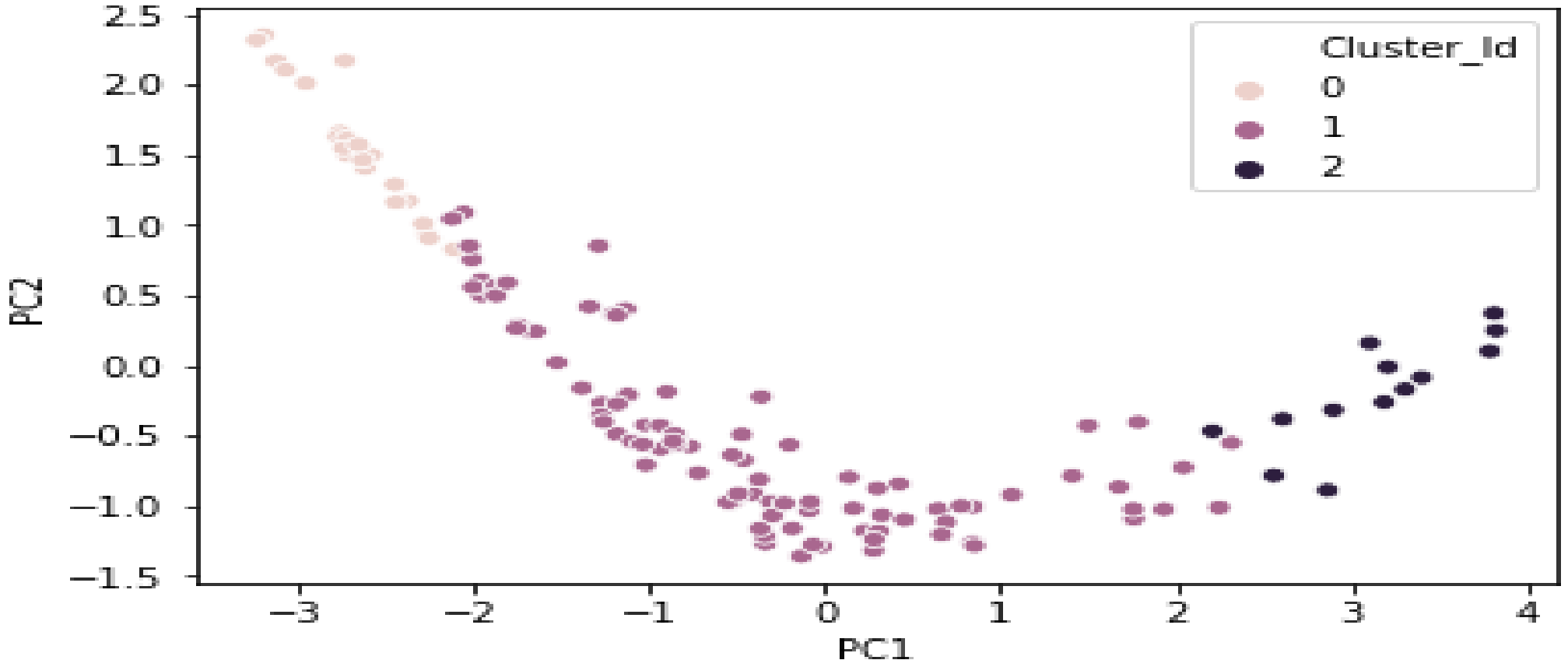
Hierarchy clustering-1

- First, dendrogram using ‘single’ linkage was plotted.
- As can be seen from the plot in the top right corner, it isn't too intuitive or clear to assess the correct number of clusters
- Next, dendrogram was plotted using ‘complete’ linkage
- As expected, dendrogram from 'complete' linkage clearly shows the linkages between the clusters and the resulting inverted tree that is formed
- Cutting the tree at point '7' resulted in clusters with longest linkages as the link 'distance' is maximum.
- The number of resulting clusters are ‘3’, same as per K-means clustering
- Cluster Ids were calculated for the three clusters using hierarchy clustering algorithm
- Next, the three clusters were formed by associating respecting Cluster Ids, first to the five PCs dataframe and then to countries by merging this dataframe with parent dataframe consisting of countries and original variables



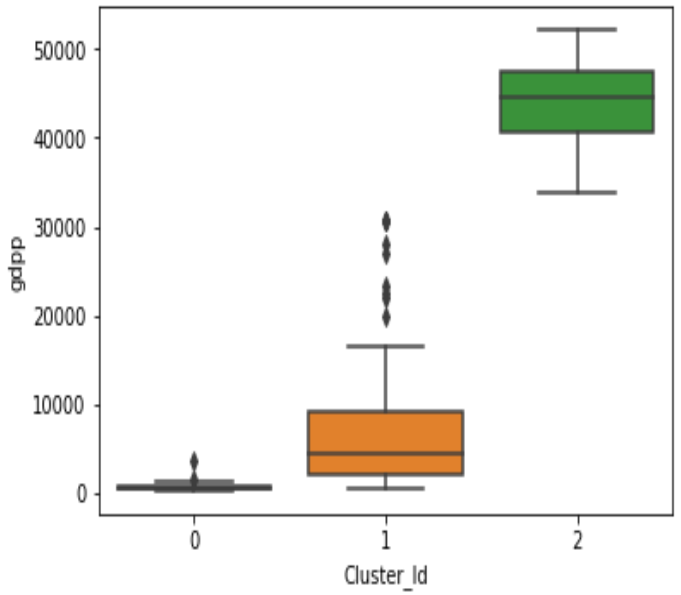


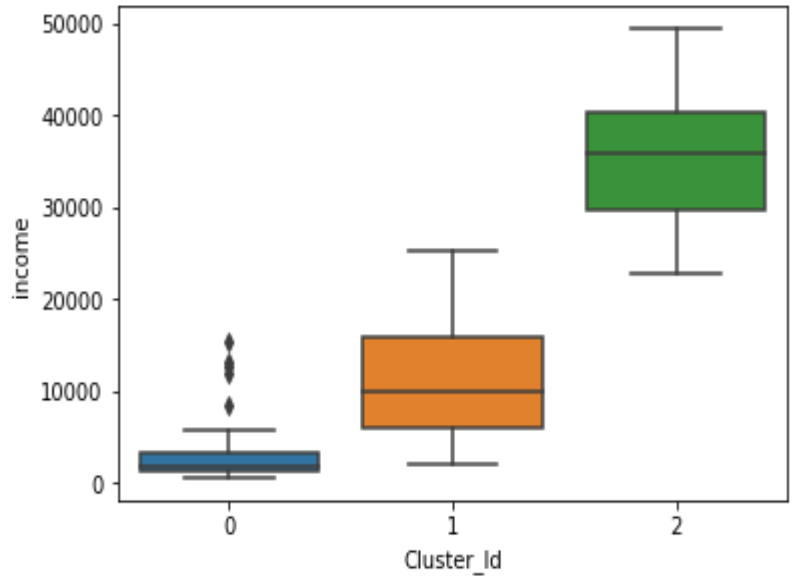
Scatterplot between PC1 and PC2, with points representing their respective clusters

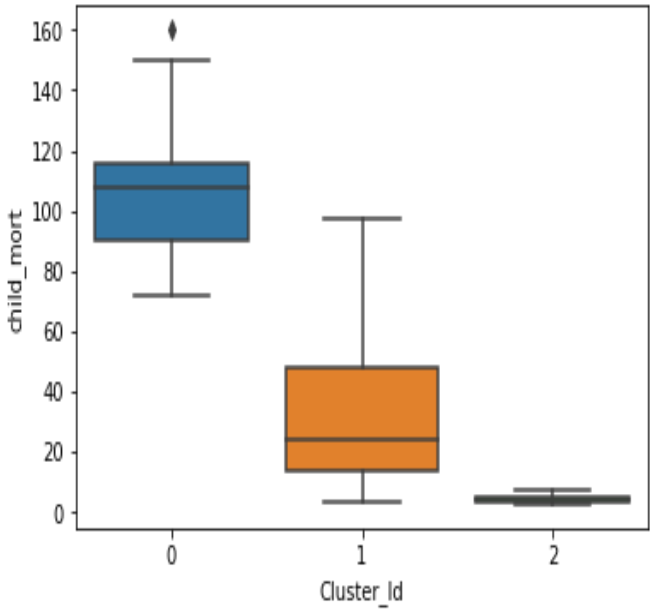


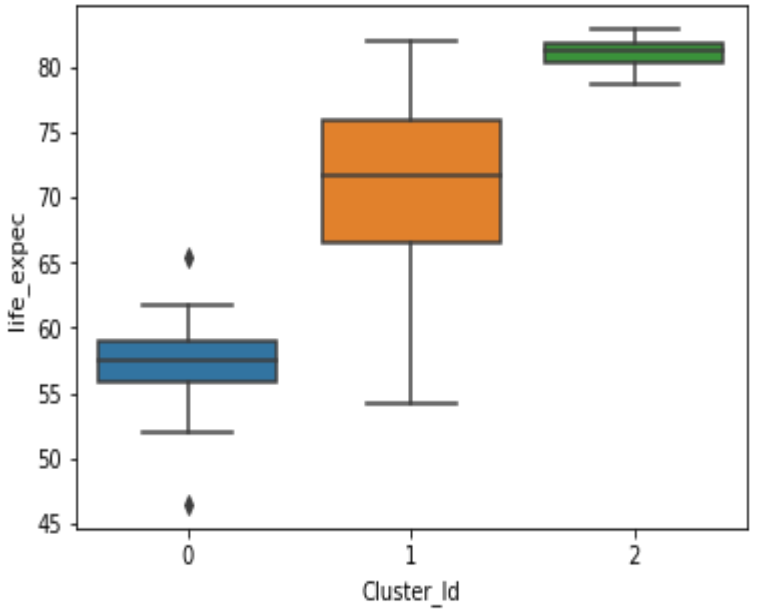
Hierarchy clustering-2: Boxplot analysis

- It can be seen from box plots that here too, cluster -0 has the countries of interest due to the following analysis results:
- The 'gdpp' is lowest for this cluster. Therefore, the spending on 'health', 'imports' and 'exports' is least and in fact GDP per capita, given by 'gdpp' is well below acceptable limits.
- The 'child_mort' rate is highest for this cluster and again, alarmingly high, so countries within this cluster need the most attention
- The 'income' is lowest for this cluster so the spending power of an individual from countries within this cluster is lowest which means that access to basic needs is toughest for people of these countries
- Therefore, created a dataframe with countries from cluster-0 only for further analysis









Hierarchy clustering-3: Shortlisting dire state countries

- First shortlisted the countries within the dataframe that have values of 'gdpp', 'income' lower than or equal to their mean , and, have 'child_mort' higher than or equal to the mean value.
- This resulted in reduced dataset but still had a number of countries which did not require immediate financial aid.
- Next, filtered this dataset based on 'total_fer' being greater than or equal to ‘5’.
- The resulting dataset consisted reduced number of countries however, further filtering was still possible
- Filtered the countries further based on 'child_mort’, as the 'income' and 'gdpp' are too low for all the countries so not fair to reduce the list of primary candidates for aid based these parameters.
- Derived the dataframe consisting of countries that are in dire need of funds.

Hierarchy clustering- 4: Final dataframe with shortlisted countries and respective parameters

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	Cluster_Id
18	Burkina Faso	116.0	110.40	38.76	170.20	1430	6.81	57.9	5.87	575	0
29	Congo, Dem. Rep.	116.0	137.27	26.42	165.66	609	20.80	57.5	6.54	334	0
52	Guinea	109.0	196.34	31.95	279.94	1190	16.10	58.0	5.34	648	0
53	Guinea-Bissau	114.0	81.50	46.50	192.54	1390	2.97	55.6	5.05	547	0
88	Niger	123.0	77.26	17.96	170.87	814	2.55	58.8	7.49	348	0
101	Sierra Leone	160.0	67.03	52.27	137.66	1220	17.20	55.0	5.20	399	0

The list of countries shortlisted via K-means clustering and Hierarchy clustering turned out to be the same in this case and the list of countries that need the funds the most are:

- Burkina Faso
- Congo, Dem. Rep.
- Guinea
- Guinea-Bissau
- Niger
- Sierra Leone