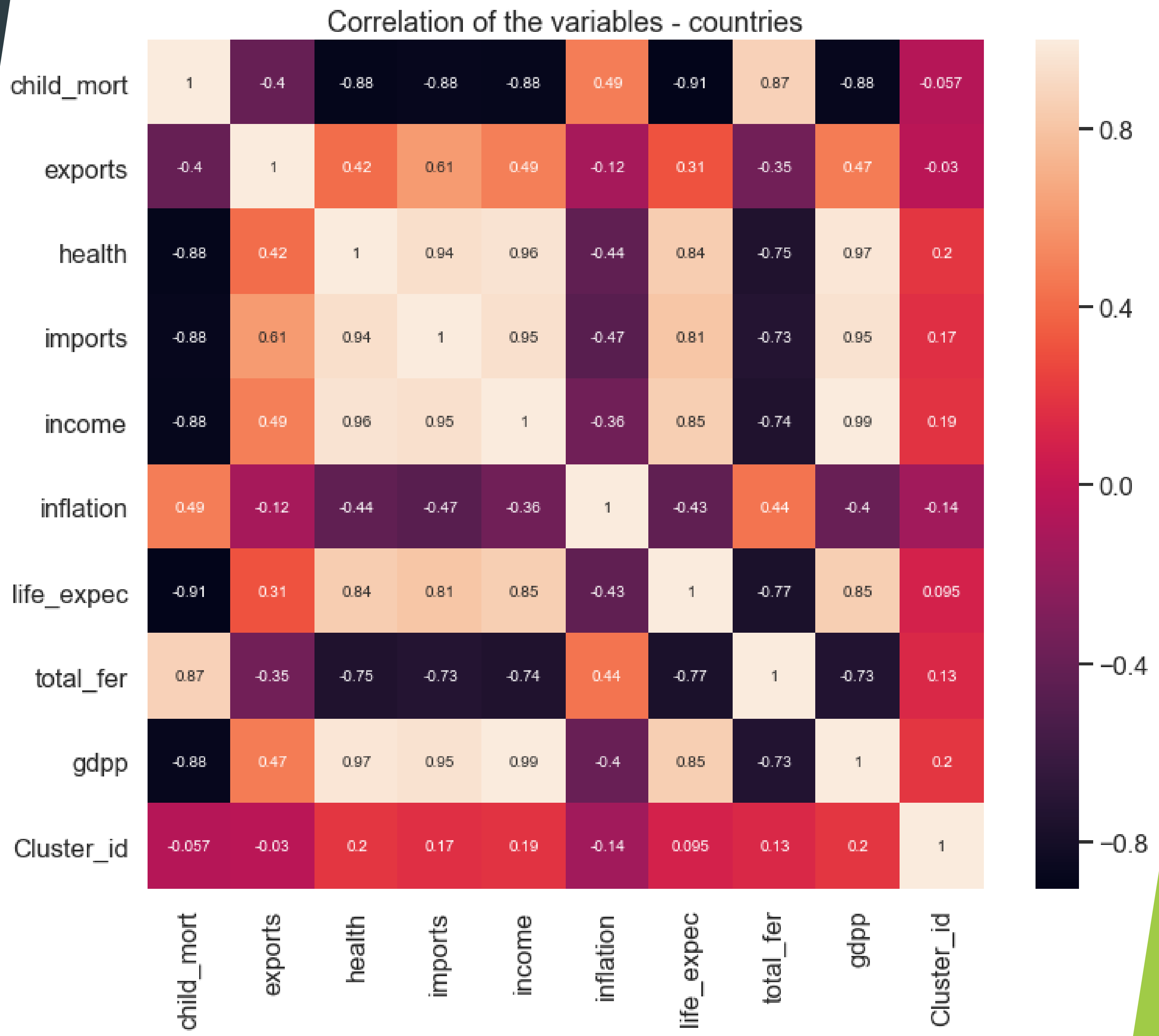
Clustering and PCA

- BY SANGRAM SINHA

Visualization of the Countries/Correlation of the variables

From the heatmap on the right we can conclude below points –

- Some of the variables are having high positive correlation like income and gdpp, health, imports etc.
- Some of them have high negative correlation like child mortality and life expectancy, gdpp, health etc.
- This shows that the dataset is having multicollinearity.



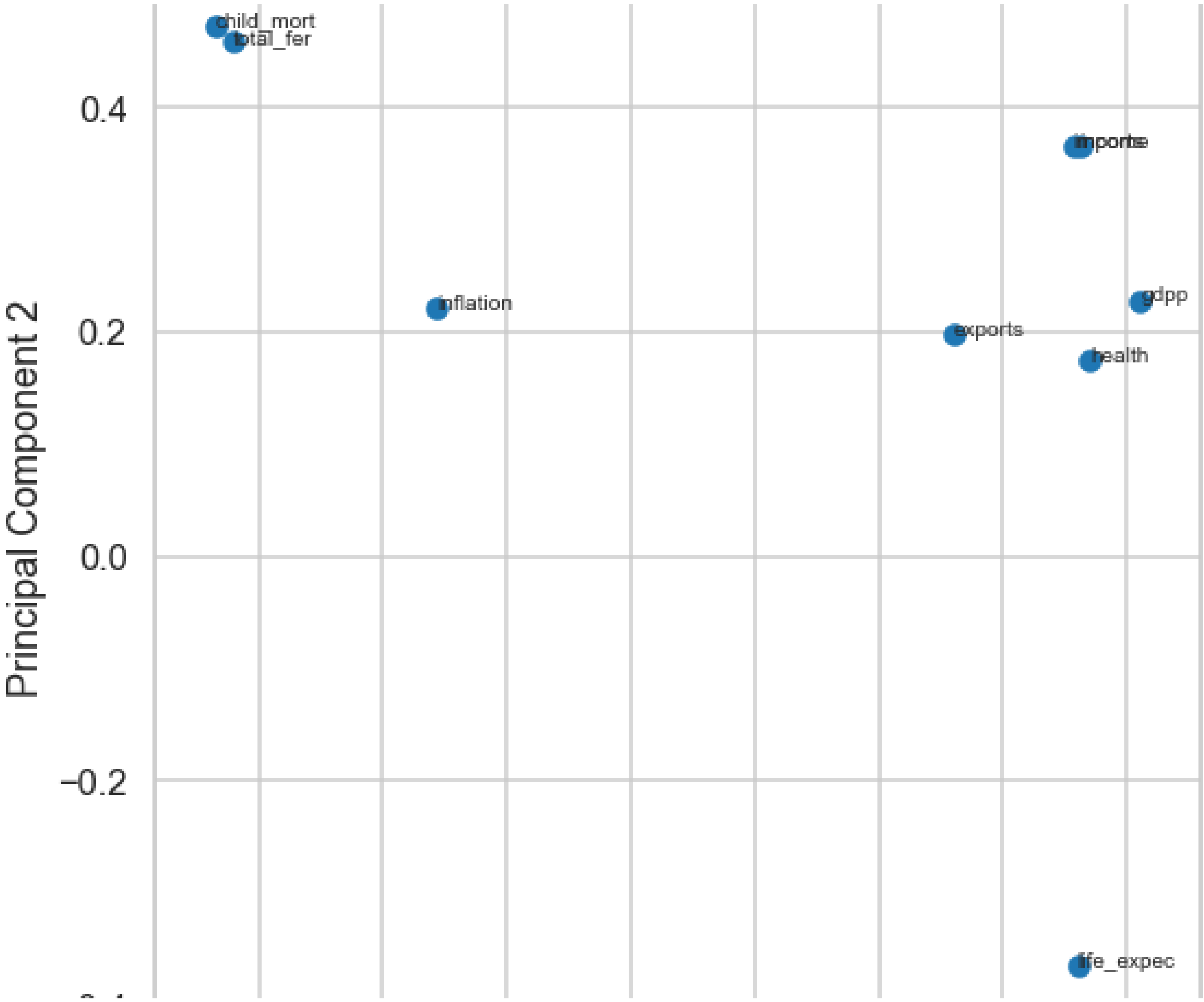
Correlation of the variables - countries

	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	Cluster_id
child_mort	1	-0.4	-0.88	-0.88	-0.88	0.49	-0.91	0.87	-0.88	-0.057
exports	-0.4	1	0.42	0.61	0.49	-0.12	0.31	-0.35	0.47	-0.03
health	-0.88	0.42	1	0.94	0.96	-0.44	0.84	-0.75	0.97	0.2
imports	-0.88	0.61	0.94	1	0.95	-0.47	0.81	-0.73	0.95	0.17
income	-0.88	0.49	0.96	0.95	1	-0.36	0.85	-0.74	0.99	0.19
inflation	0.49	-0.12	-0.44	-0.47	-0.36	1	-0.43	0.44	-0.4	-0.14
life_expec	-0.91	0.31	0.84	0.81	0.85	-0.43	1	-0.77	0.85	0.095
total_fer	0.87	-0.35	-0.75	-0.73	-0.74	0.44	-0.77	1	-0.73	0.13
gdpp	-0.88	0.47	0.97	0.95	0.99	-0.4	0.85	-0.73	1	0.2
Cluster_id	-0.057	-0.03	0.2	0.17	0.19	-0.14	0.095	0.13	0.2	1

PCA components(PC1 vs PC2)

From the plot on the left we can conclude below points.

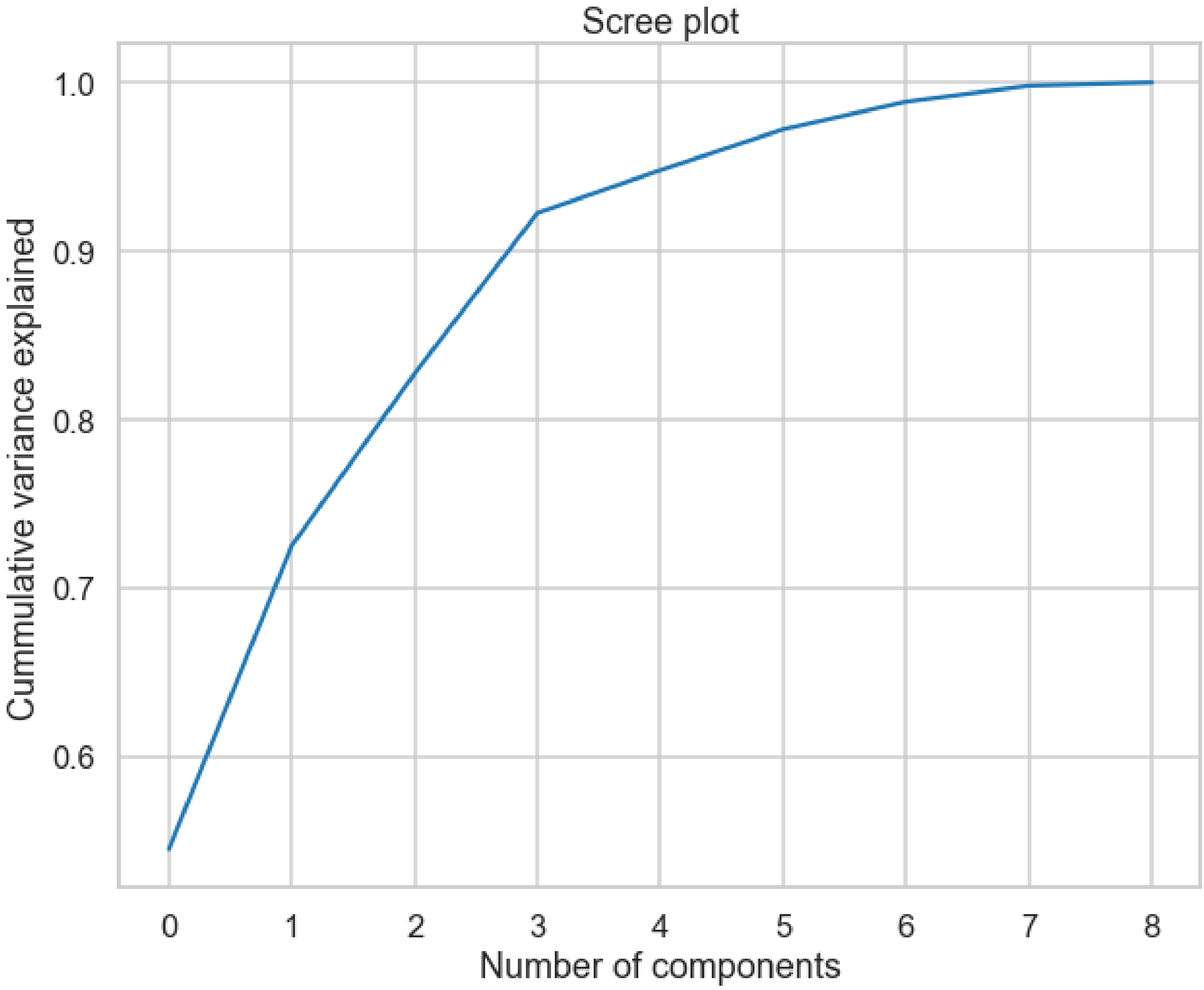
- The data points of life expectancy is in the direction of principal component 1.
- All the high value data points are in the direction of principal component 1.



PCA - Scree plot

From the graph on the left points to be concluded –

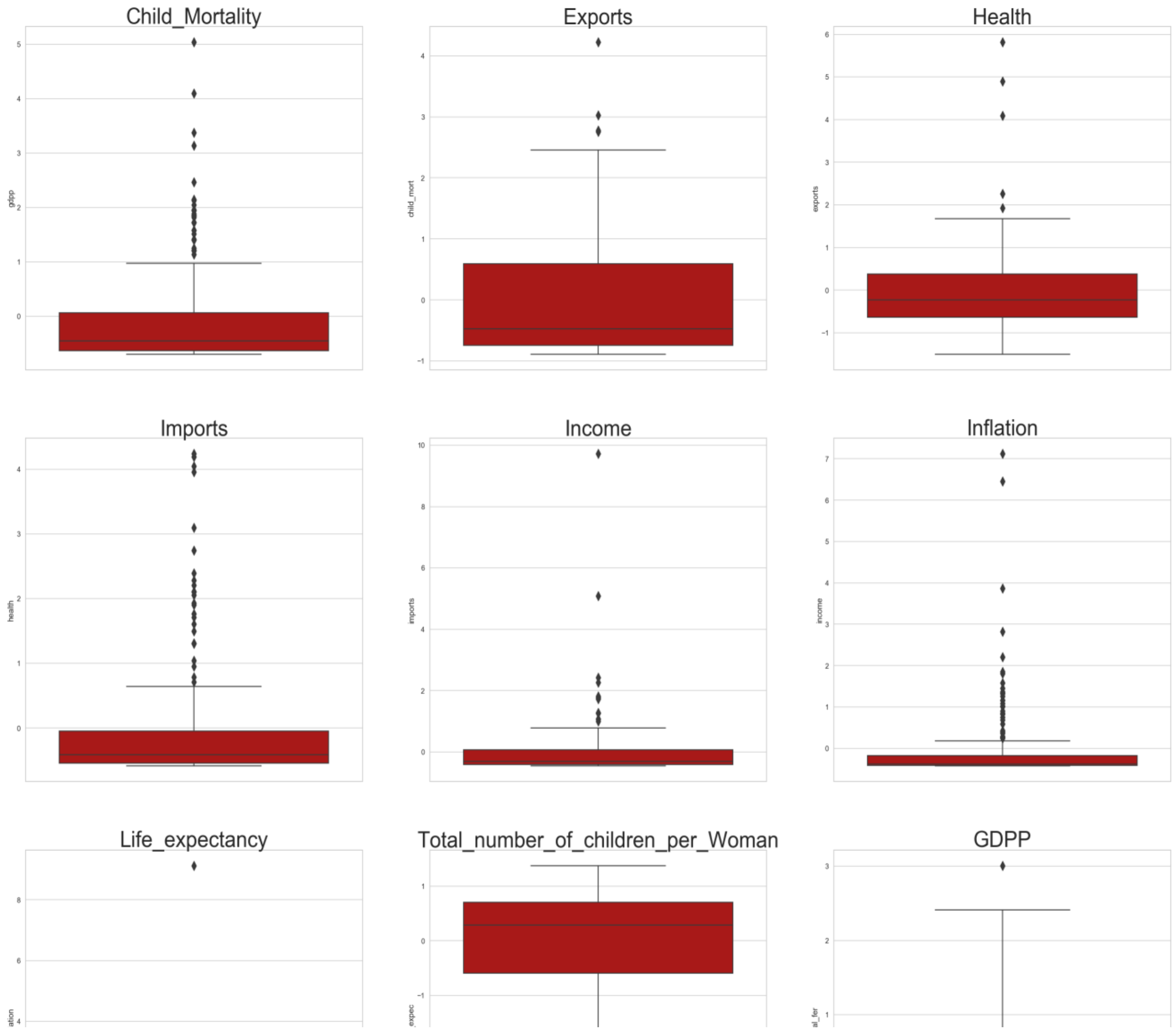
- The number of component equal to 4 is having approx. 95% of variance explained.
- The number of component equal to 5 is having approx. 97% of variance explained.
- So, the ideal number of components can be chosen is 5.



Visualization of Outliers

From the boxplots attached on the left, points to be concluded -

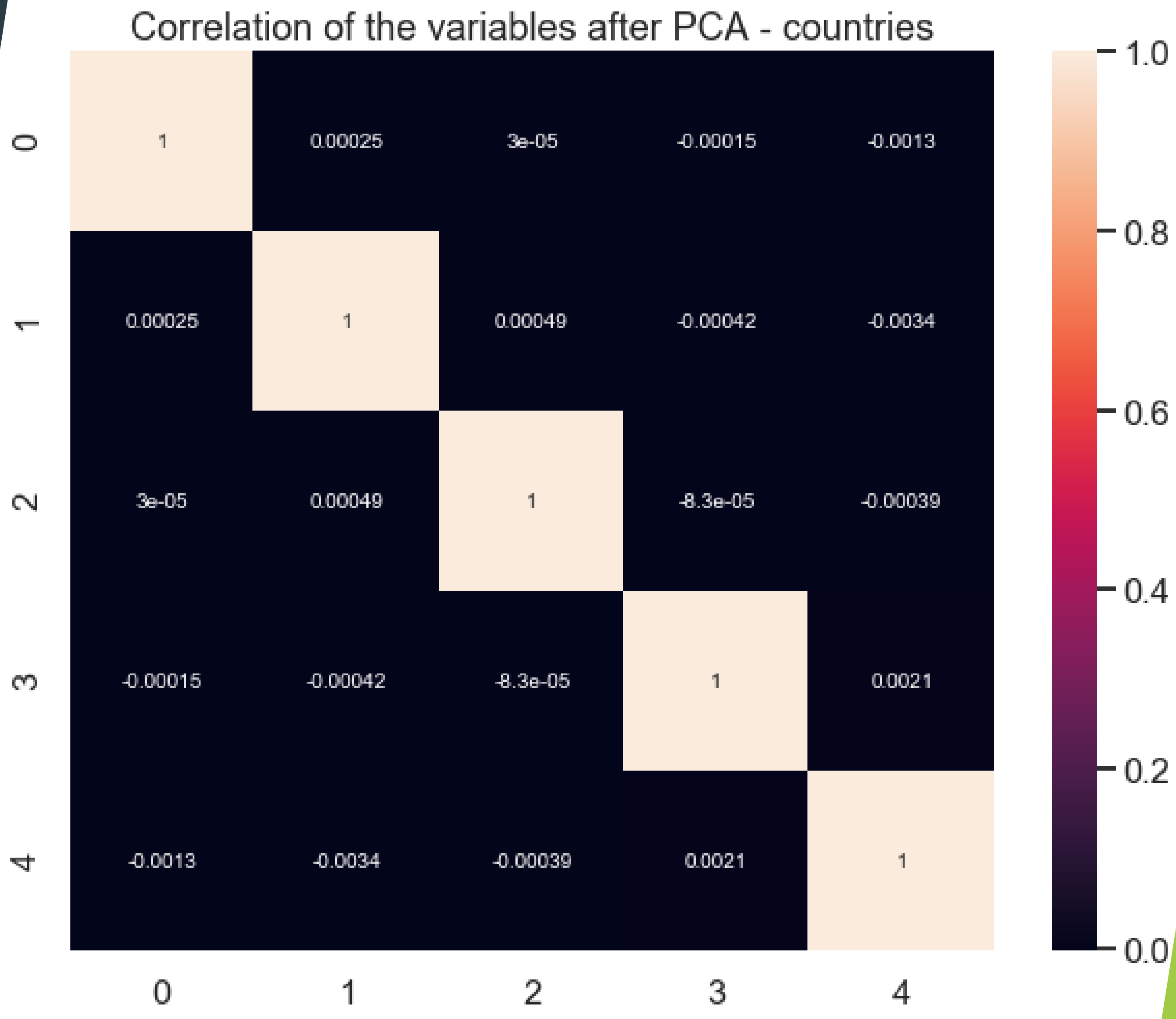
- As we can see that all the boxplots created for the variables are having decent amount of outliers
- All boxplots except one ‘Total_number_of_children_per_woman’ is having outliers on the bottom of the boxplots which means there are some countries were no. of children per woman is very less compared to the other countries.
- The Inflation boxplot is having very thin size of quartiles compared to others.



Correlation after PCA

From the graph on the left side, points to be concluded –

- All the correlation are showing in dark color, which means they all are close to 0.
- This shows that after doing PCA we have removed the multicollinearity.



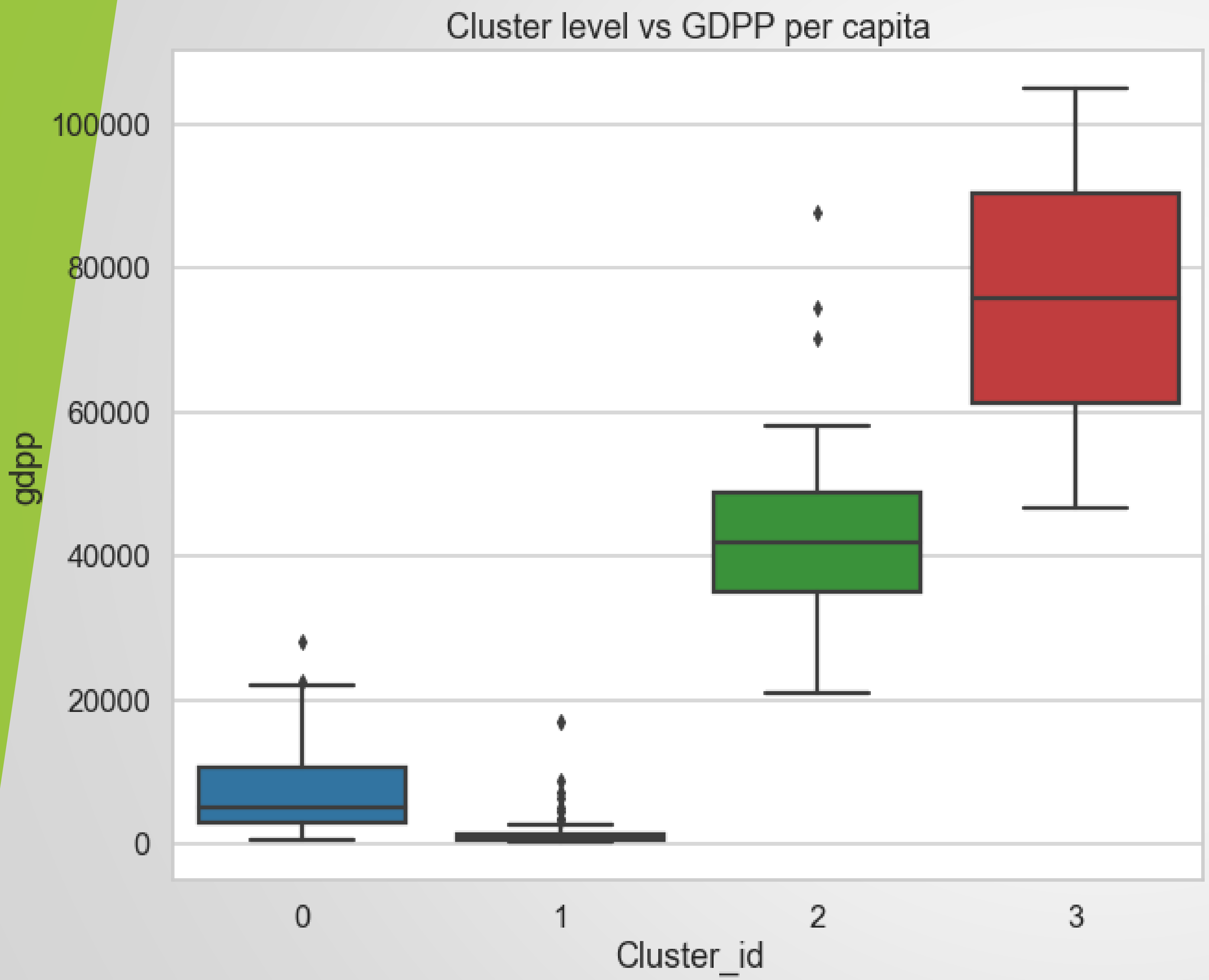
Correlation of the variables after PCA - countries

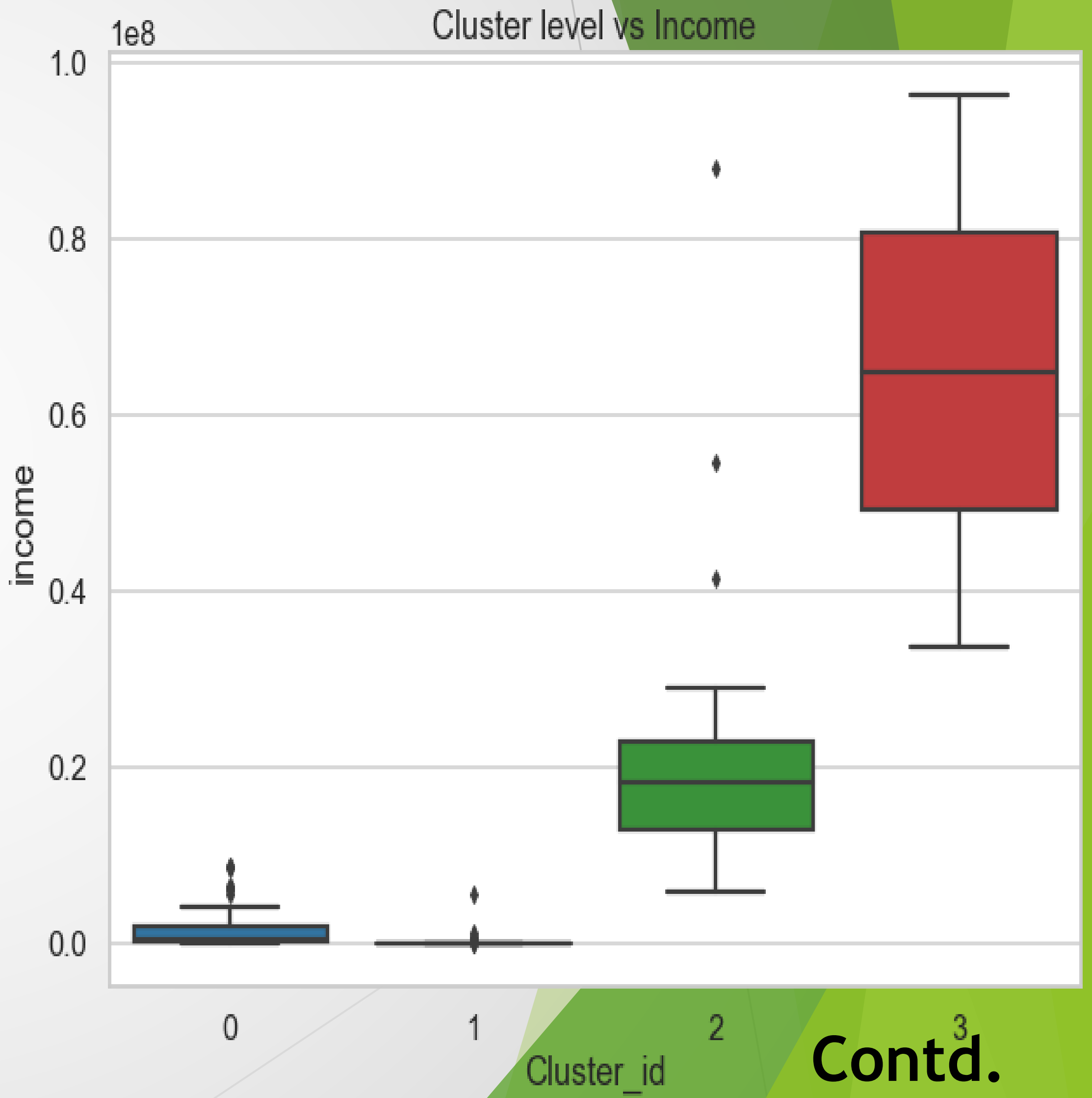
	0	1	2	3	4
0	1	0.00025	3e-05	-0.00015	-0.0013
1	0.00025	1	0.00049	-0.00042	-0.0034
2	3e-05	0.00049	1	-8.3e-05	-0.00039
3	-0.00015	-0.00042	-8.3e-05	1	0.0021
4	-0.0013	-0.0034	-0.00039	0.0021	1

K-Means analysis

- We are trying to find the optimum value of the k-value based on the business requirements.
- So, to achieve this we used silhouette analysis to find the score of range of cluster values.
- We found below silhouette scores:
  - For no. of cluster=2,silhouette score is 0.436593354493332
  - For no. of cluster=3,silhouette score is 0.4111424002436615
  - For no. of cluster=4,silhouette score is 0.4166607876891153
  - For no. of cluster=5,silhouette score is 0.42468902258978025
  - For no. of cluster=6,silhouette score is 0.3016100082966709
  - For no. of cluster=7,silhouette score is 0.30904997348203855
  - For no. of cluster=8,silhouette score is 0.3052243161399318
  - For no. of cluster=9,silhouette score is 0.28967937743824324
- We found that cluster =2 is having highest score hence k value should be 2 but, if we choose k value as '2' , it will not suit business needs.
- Hence, we use k value as '4' since it is giving precise information also fulfills business needs.

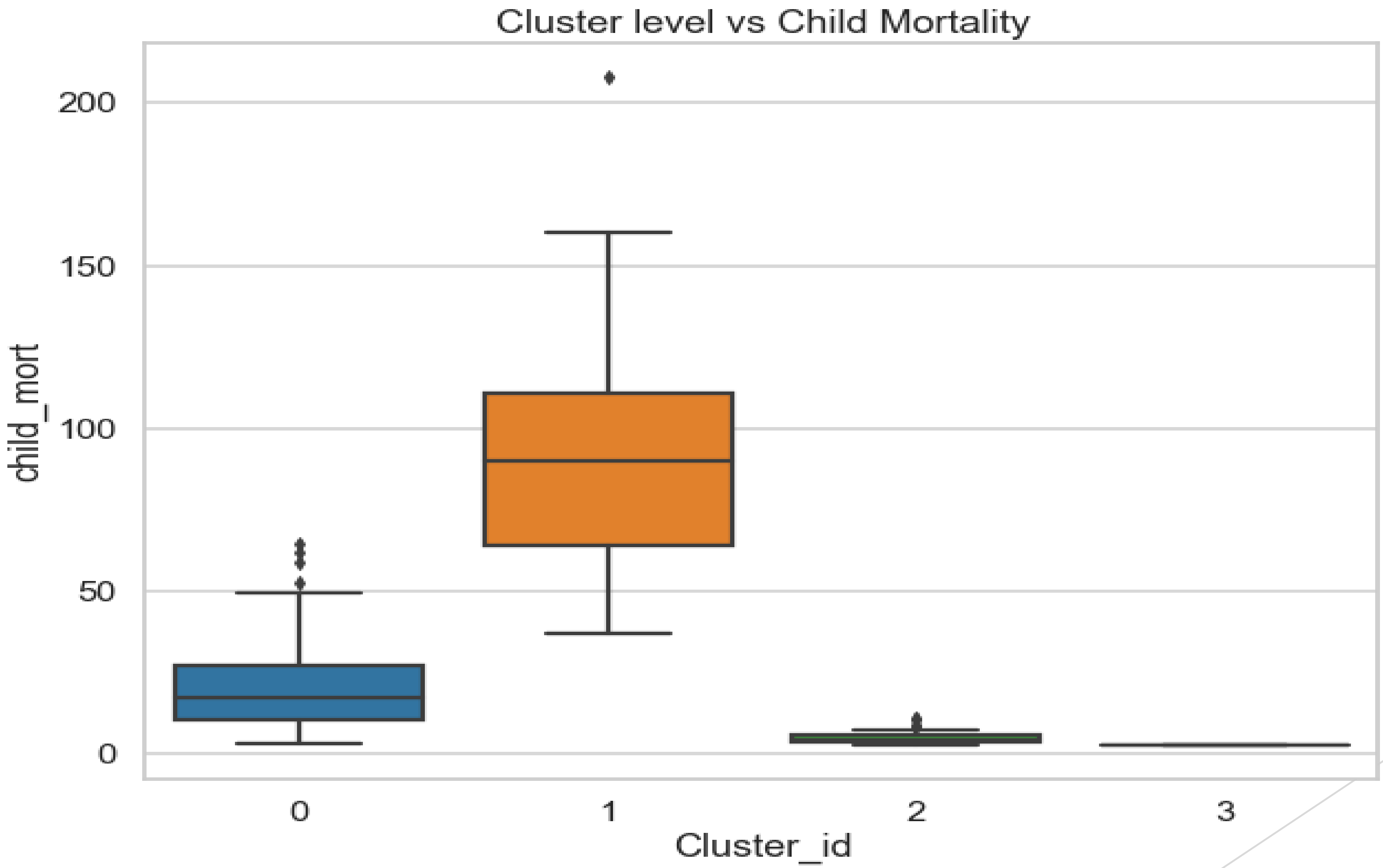
Visualization of the original variables with clusters





Contd.

Visualization of the original variables with clusters



We are using **cluster** = 4 as it fits perfectly in business sense, if you could see the boxplots from the previous two slides, they are giving promising output w.r.t original variables

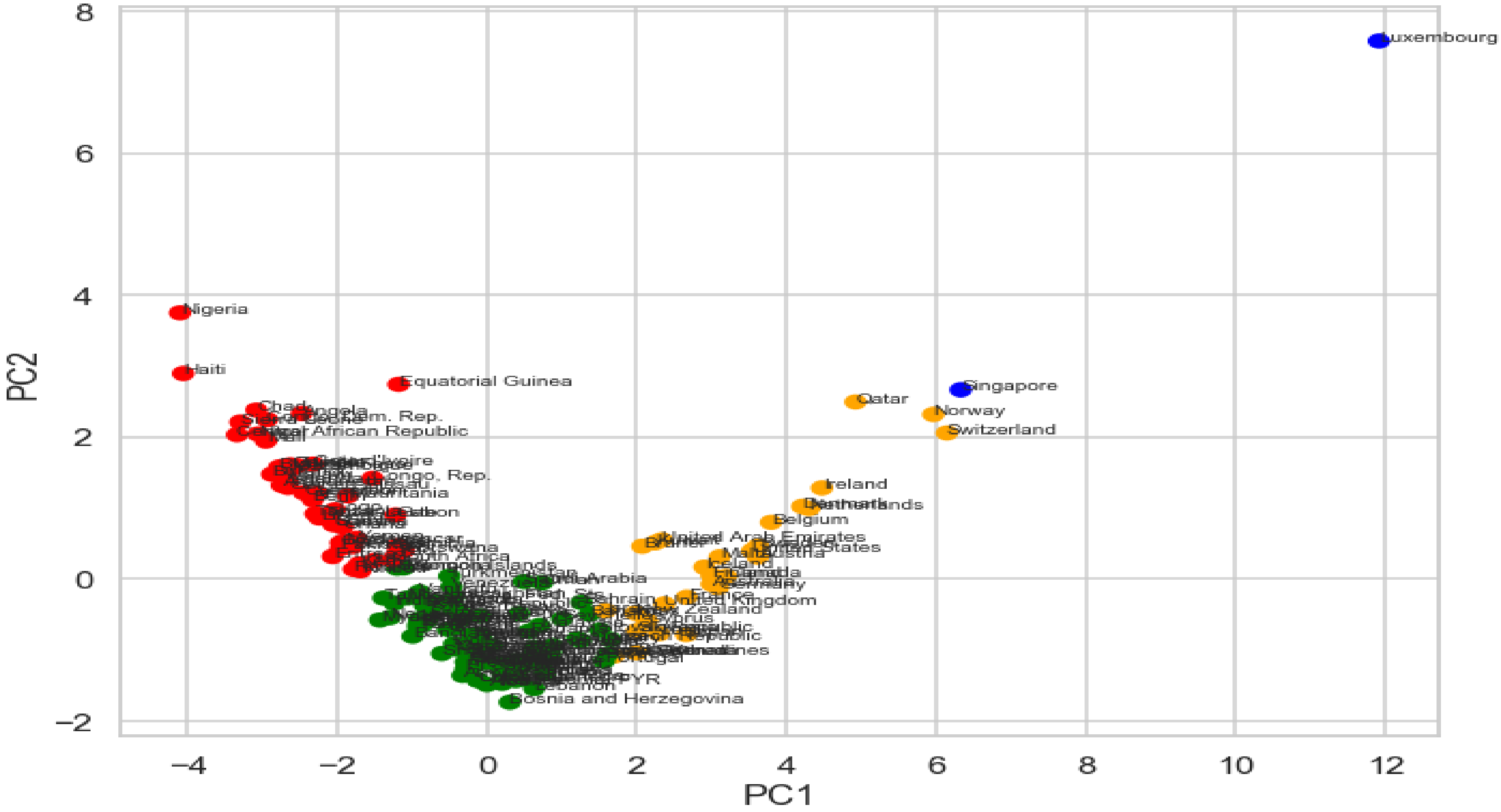
Valuable Insights from boxplot shown on the previous two slide – gdpp, income and chid_mortality.

For **cluster 0** : gdpp and income are slightly higher than the lowest of gdpp and income, Child Mortality is also little higher with outliers

For **cluster 1** : gdpp and income is the lowest than other clusters, Mortality of children is very high than other clusters.

For **cluster 2** : Behaving normally in all departments(income, gdpp and children mortality) except for some outliers.

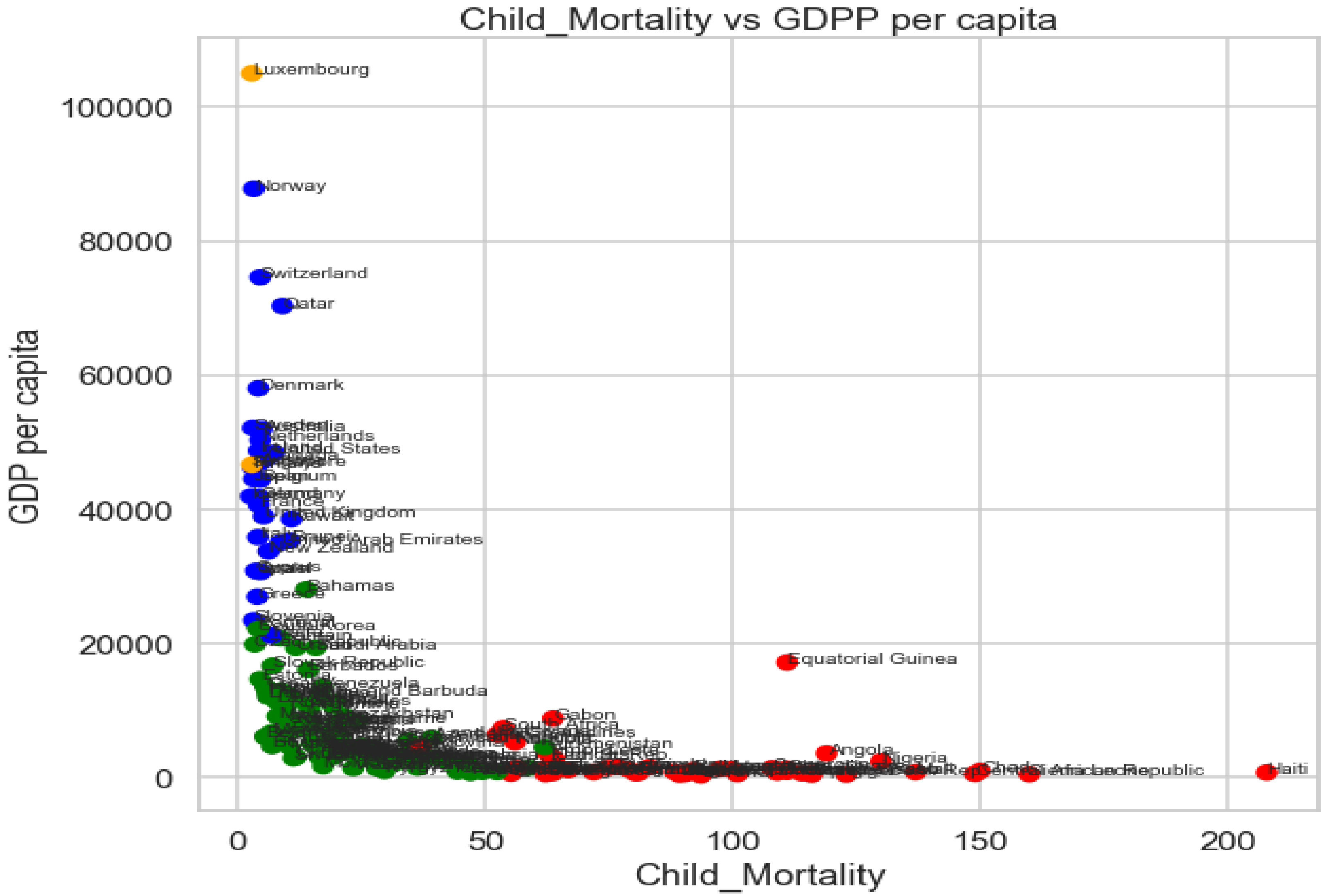
For **cluster 3** : gdpp and income is higher than other clusters, Mortality of children is very less compared to other clusters.



Visualization of principal component 1 and 2 in X-Y axes

Insights drawn from the scatter plot visualization

- As we can see from first two principal components(PC1 & PC2), the PC1 is in the direction where the countries need of least help. Here, why we are choosing PC1 because it has maximum percentage of variance explained.
- We can see that countries like ‘Singapore’ and ‘Luxembourg’ are having high PC1 which means they are doing well, where on other hand countries like ‘Nigeria’, ‘Hiti’, ‘Equatorial Guniea’ and close to them requires urgent need of aid.
- Hence from the datapoints/countries of color ‘Red’ are in direst need of aid than the color in 'Orange'



Visualizing with two original variables (Child_mort vs gdpp)

Insights drawn from the scatter plot visualization

Three points we can conclude :

- Country ‘Haiti’ is in dire need of aid
- Country ‘Luxembourg’ is having good gdpp and less child mortality rate
- ‘Haiti’ and ‘Luxembourg’ are two outliers we can see from the scatter plot

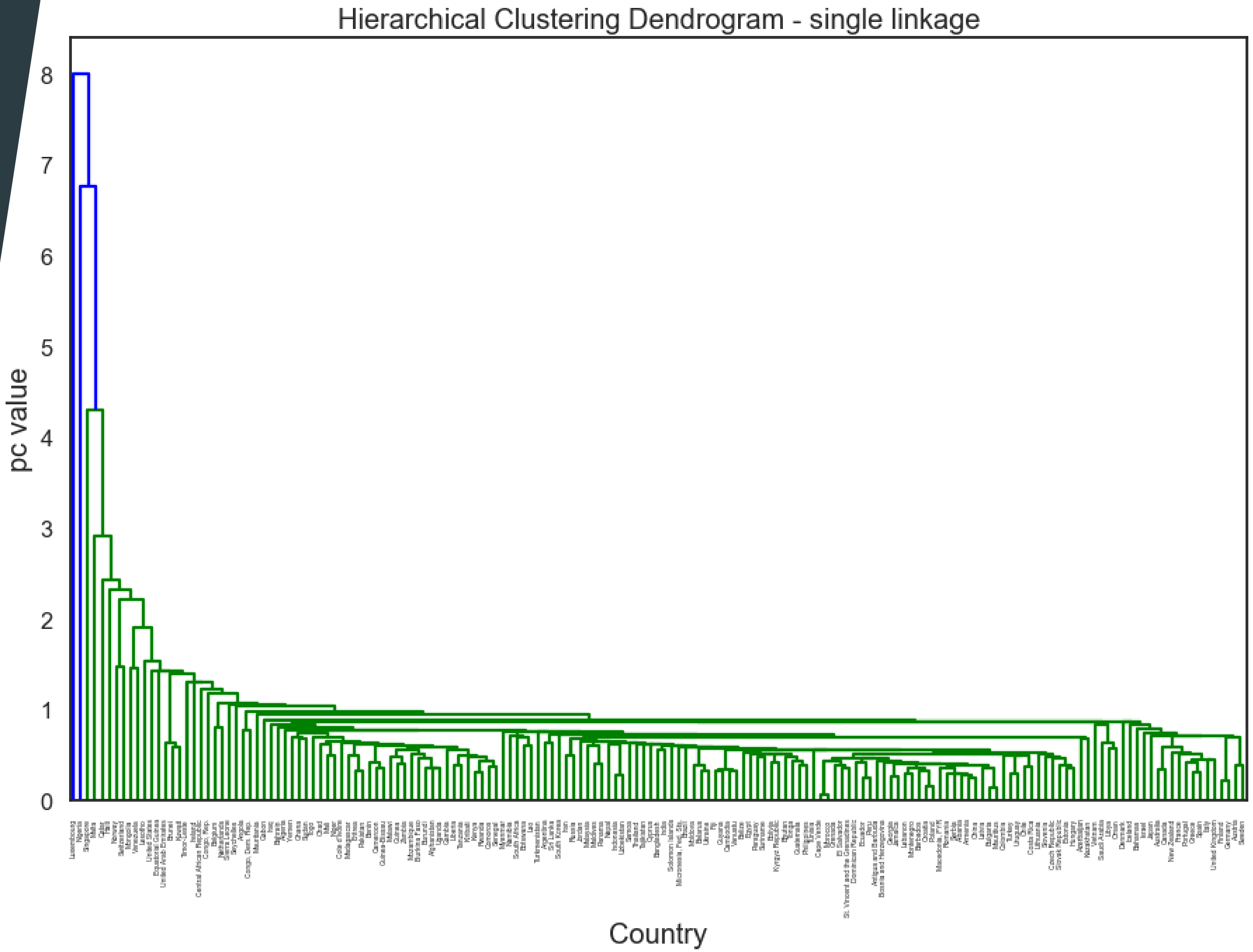
Insights (K-Means with outliers)

- There are total 47 countries from the dataset need urgent help/aid as they are having lowest income, high child mortality and low gdp per capita.
- Only 2 countries is there with good socio-economic and health factors

Hierarchical clustering (Linkage)

This is another method to find our low development countries.

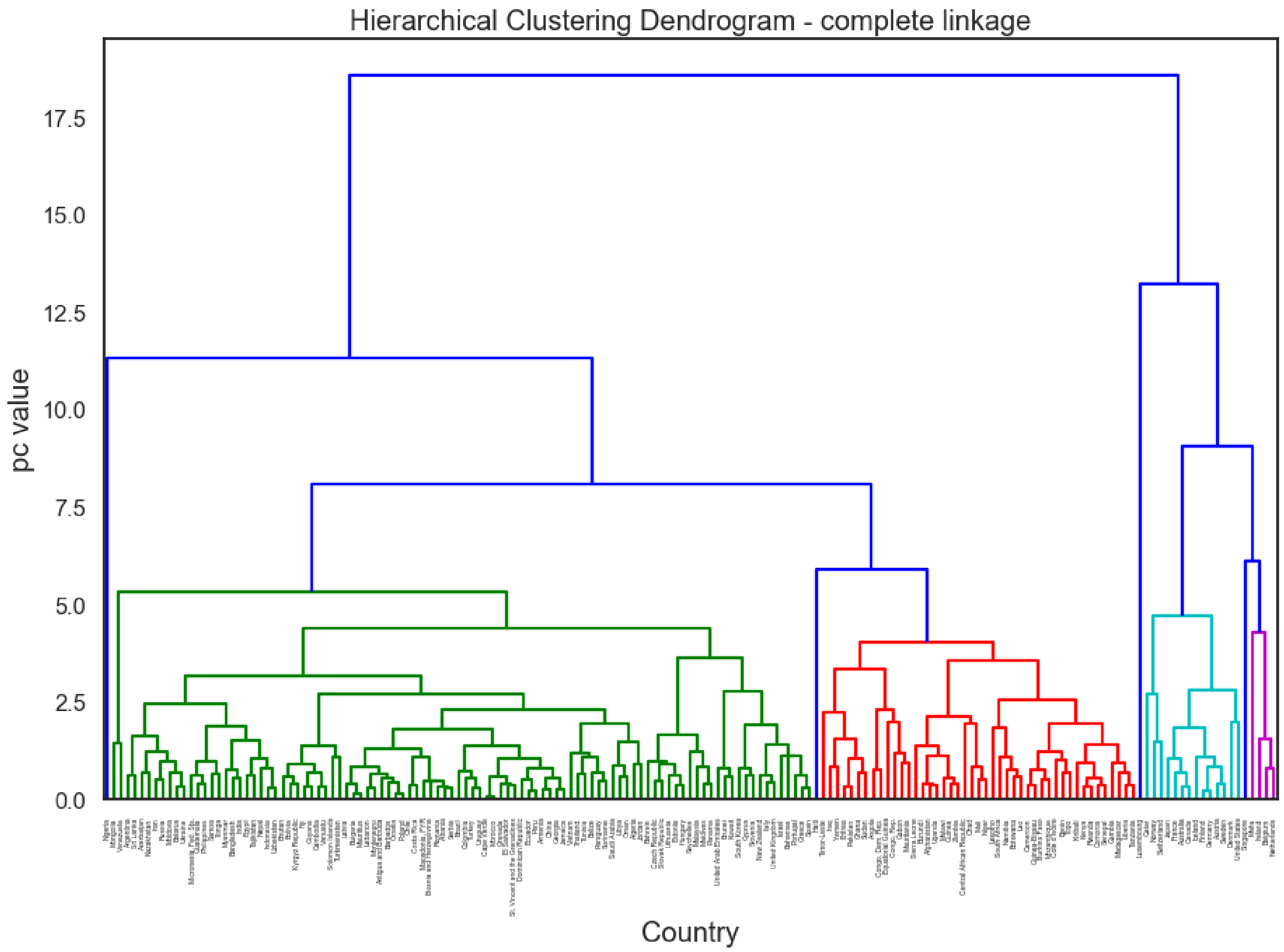
As we can see from the graph of linkage dendrogram, it is not quite visible and doesn’t not suits properly with the dataset because we can cut the tree in a threshold value, we will use complete linkage dendrogram for hierarchical clustering.



Hierarchical clustering (Complete Linkage)

Points noted from the graph on the right -

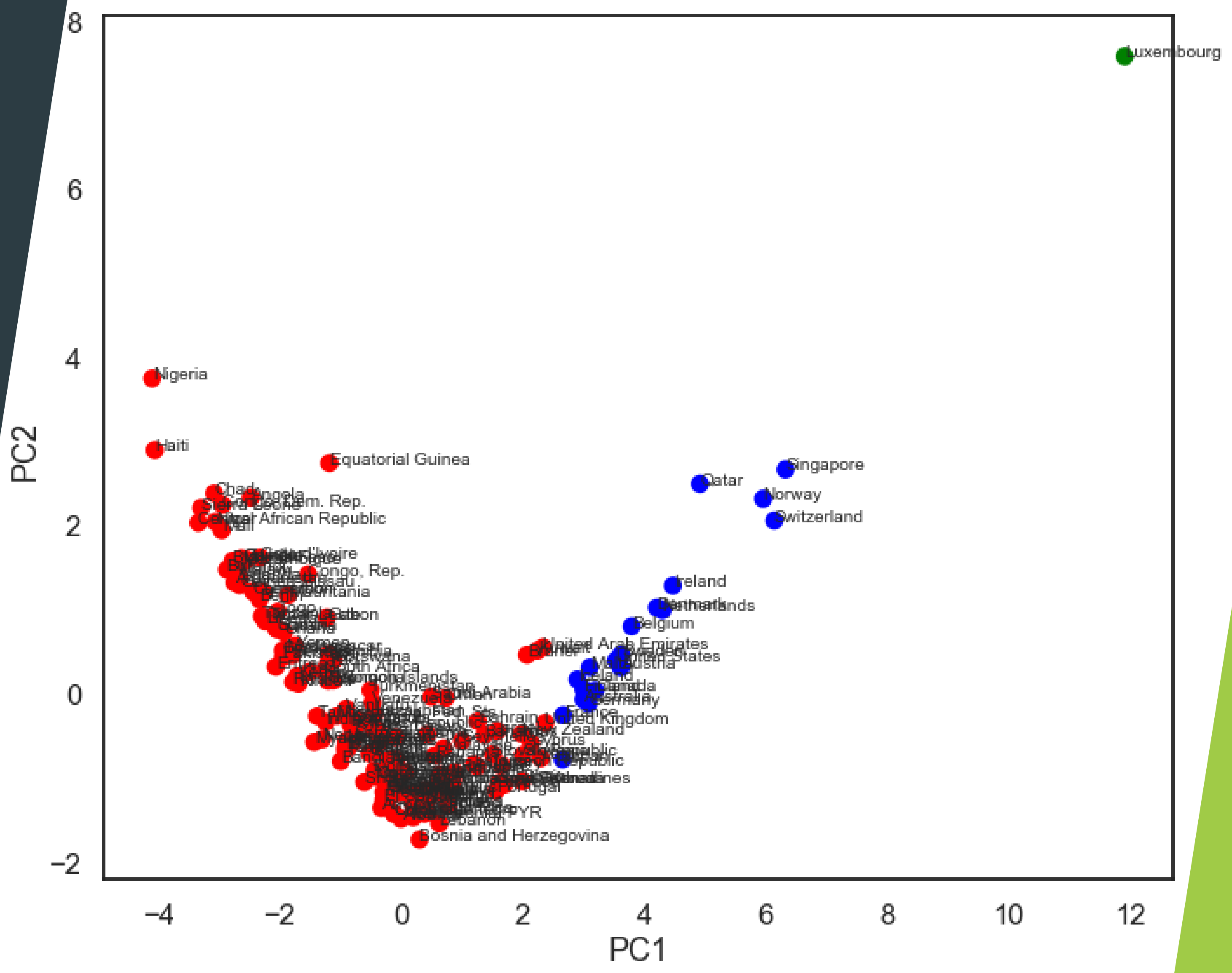
- This graph shows proper way to decide number of clusters needs to be used by cutting at threshold value.
- We will cut at 3 branches which will give us 3 clusters



Visualization of hierarchical clustering with first two principal components

Points to be drawn from the hierarchical scatter plot -

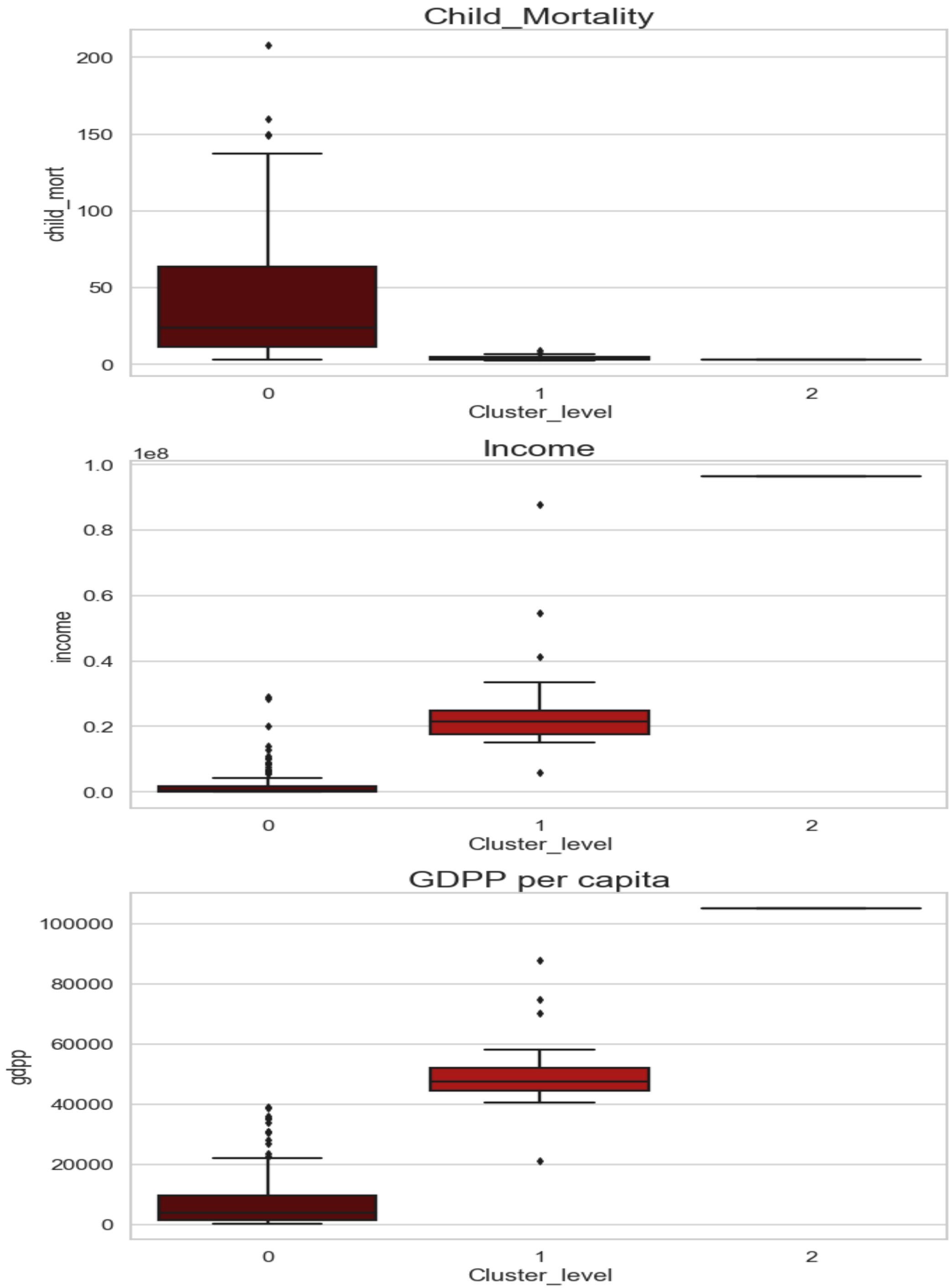
- As we can see from first two principal components(PC1 & PC2), the PC1 is in the direction where the countries need of least help. Here, why we are choosing PC1 because it has maximum percentage of variance explained.
- The ‘Red’ color datapoints of countries need help in aid but the 'Blue' one not required.



Visualization of original variables(Child mortality, Income and Gdpp)

Valuable Insights from above three boxplots -

- For cluster 0: gdpp and income is the lowest than other clusters, Mortality of children is very high than other clusters.
- For cluster 1: Behaving normally in all departments(income, gdpp and children mortality) except for some outliers.
- For cluster 2: gdpp and income is higher than other clusters, Mortality of children is very less compared to other clusters.



Insights(Hierarchical with outliers)

There are 147 countries found from the hierarchical analysis in need of urgent help/aid as it is having lowest income, high child mortality and low gdp per capita.

Conclusion - With outlier

K-Means vs Hierarchical Clustering

K-means clustering :

Countries that are direst need of aid

- Total 47 countries are in this category
- Countries that are having good socio-economic and health factors
- Total 2 countries are in this category - Luxembourg and Singapore

Hierarchical clustering :

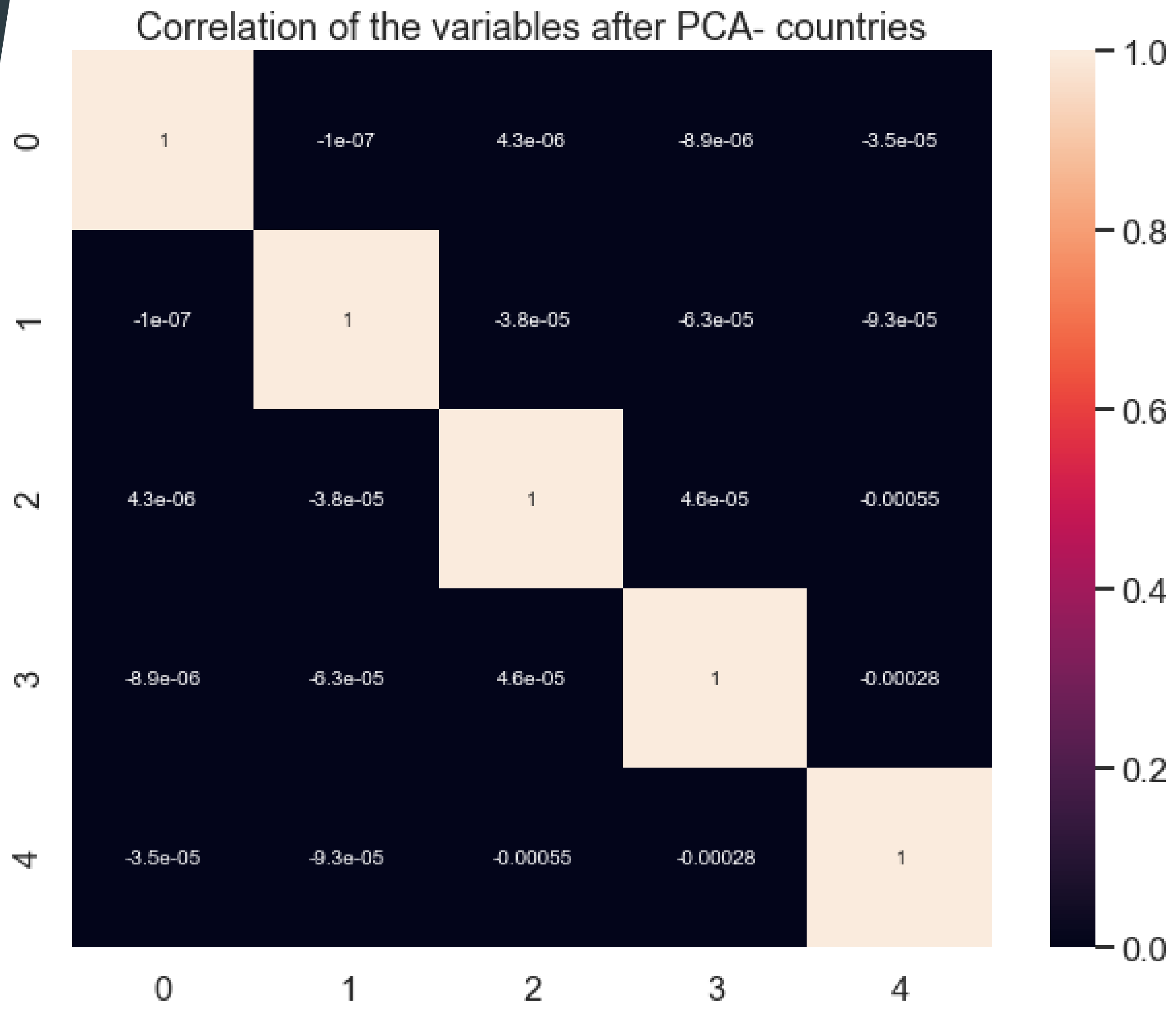
Countries that are direst need of aid

- Total 147 countries are in this category
- Countries that are having good socio-economic and health factors
- 1 country is in this category – Luxembourg
- We have seen from both methods - (K-Means and Hierarchical clustering) that extra 99 countries are being selected from hierarchical clustering. I would choose the final countries from k-means clustering as it gave **accurate** output than hierarchical clustering. I have compared the clusters and visualized from both methods and K-means gave **precise** information than hierarchical clustering.

Exclude outliers - K-Means clustering using PCA

From the graph on the left side, points to be concluded –

- All the correlation are showing in dark color, which means they all are close to 0.
- This shows that after doing PCA we have removed the multicollinearity.

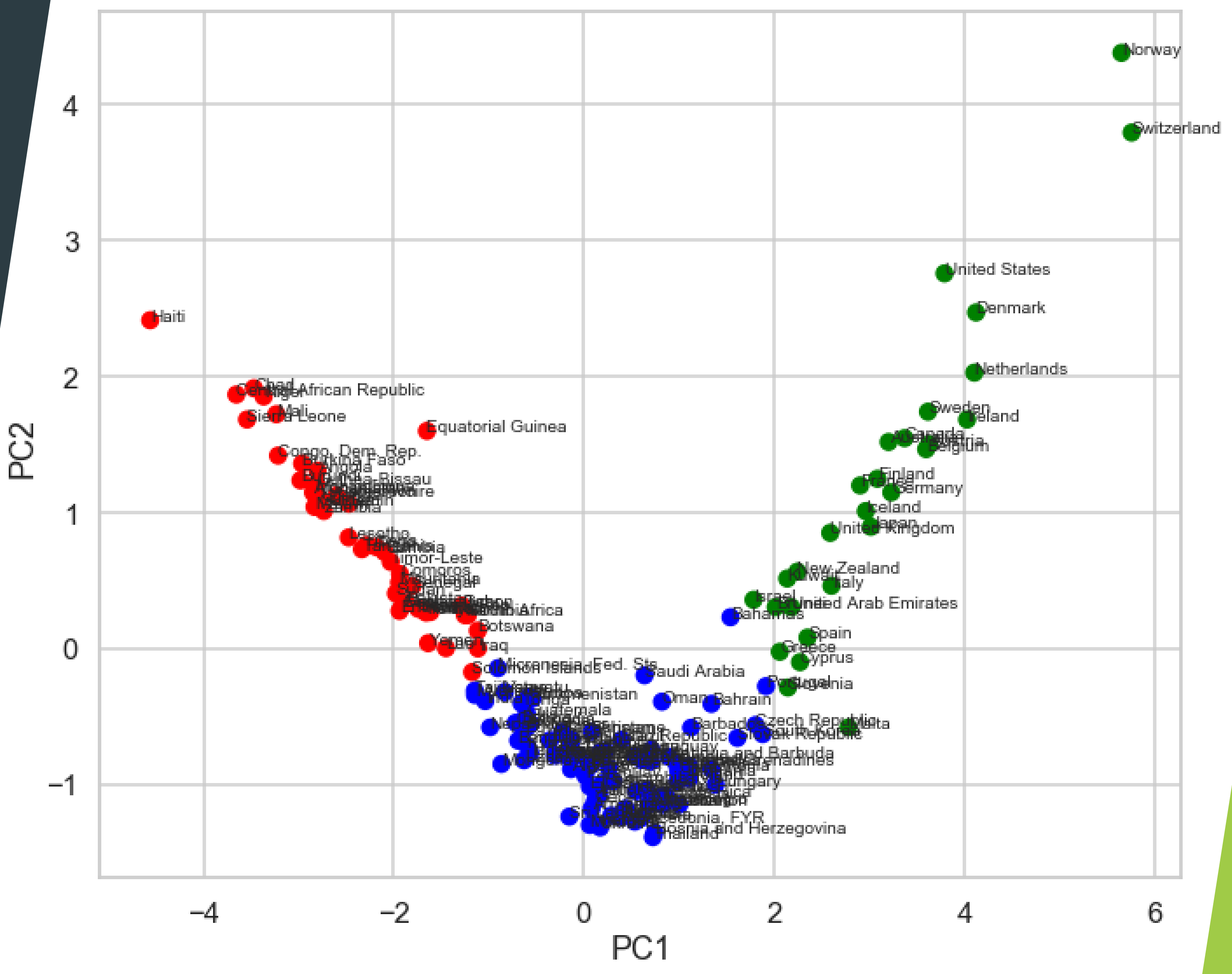


Correlation of the variables after PCA- countries

	0	1	2	3	4
0	1	-1e-07	4.3e-06	-8.9e-06	-3.5e-05
1	-1e-07	1	-3.8e-05	-6.3e-05	-9.3e-05
2	4.3e-06	-3.8e-05	1	4.6e-05	-0.00055
3	-8.9e-06	-6.3e-05	4.6e-05	1	-0.00028
4	-3.5e-05	-9.3e-05	-0.00055	-0.00028	1

Visualization with PC1 and PC2

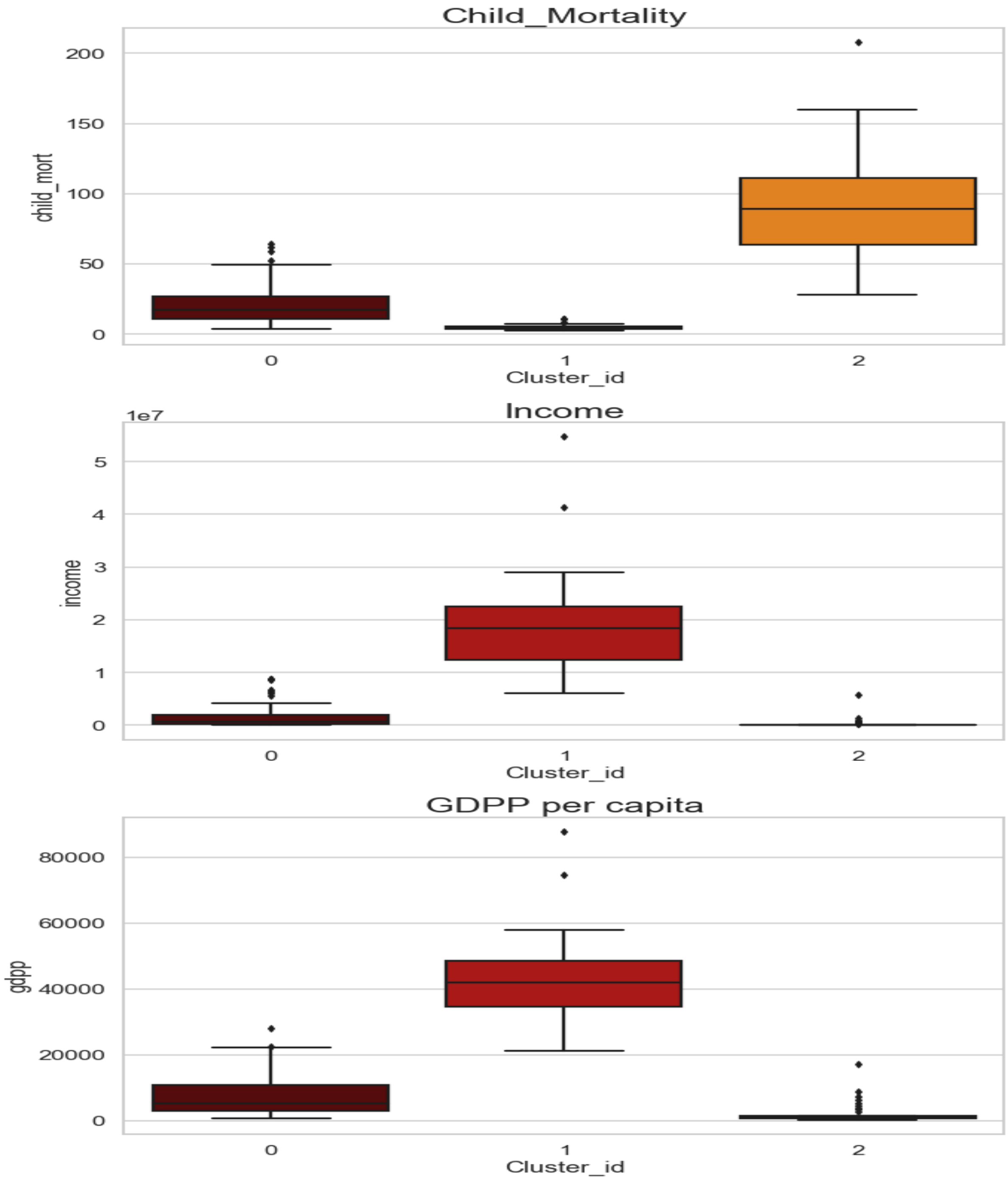
- As we can see from first two principal components(PC1 & PC2), the PC1 is in the direction where the countries need of least help. Here, why we are choosing PC1 because it has maximum percentage of variance explained.
- The Red color datapoints of countries need urgent help in aid but the 'Blue' one not required.



Visualization with original variables(gdpp, income and child mortality)

Valuable Insights from three boxplots :

- For cluster 0: Having little higher gdpp and income than cluster 1 and child mortality also acts same.
- For cluster 1: gdpp and income is higher than other clusters, Mortality of children is very less compared to other clusters.
- For cluster 2: gdpp and income is the lowest than other clusters, Mortality of children is very high than other clusters.

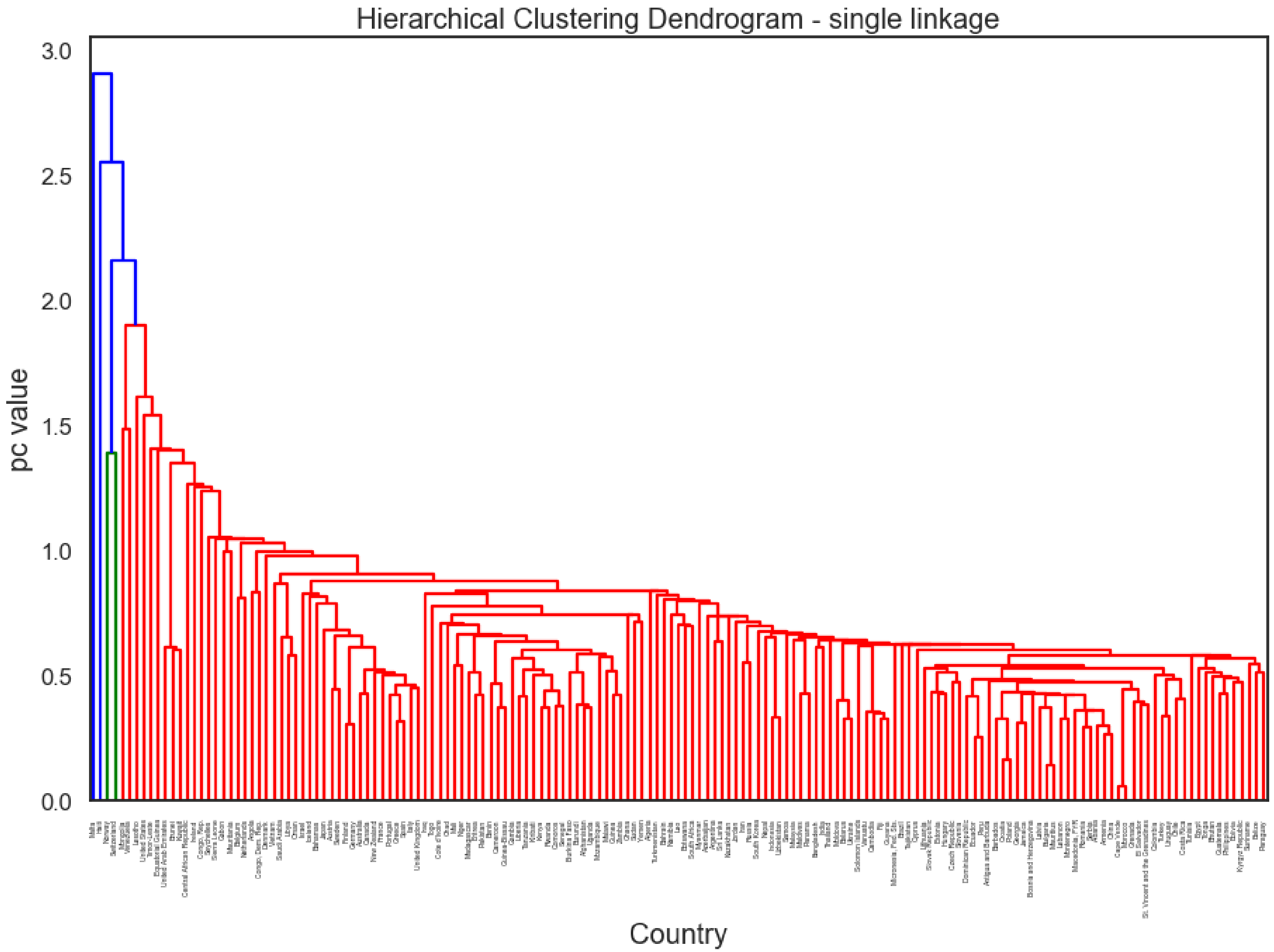


Insights (K-Means exclude outlier)

- There are total 47 countries from the dataset need of urgent help/aid as they are having lowest income, high child mortality and low gdp per capita.
- There are 28 countries having good socio-economic and health factors.

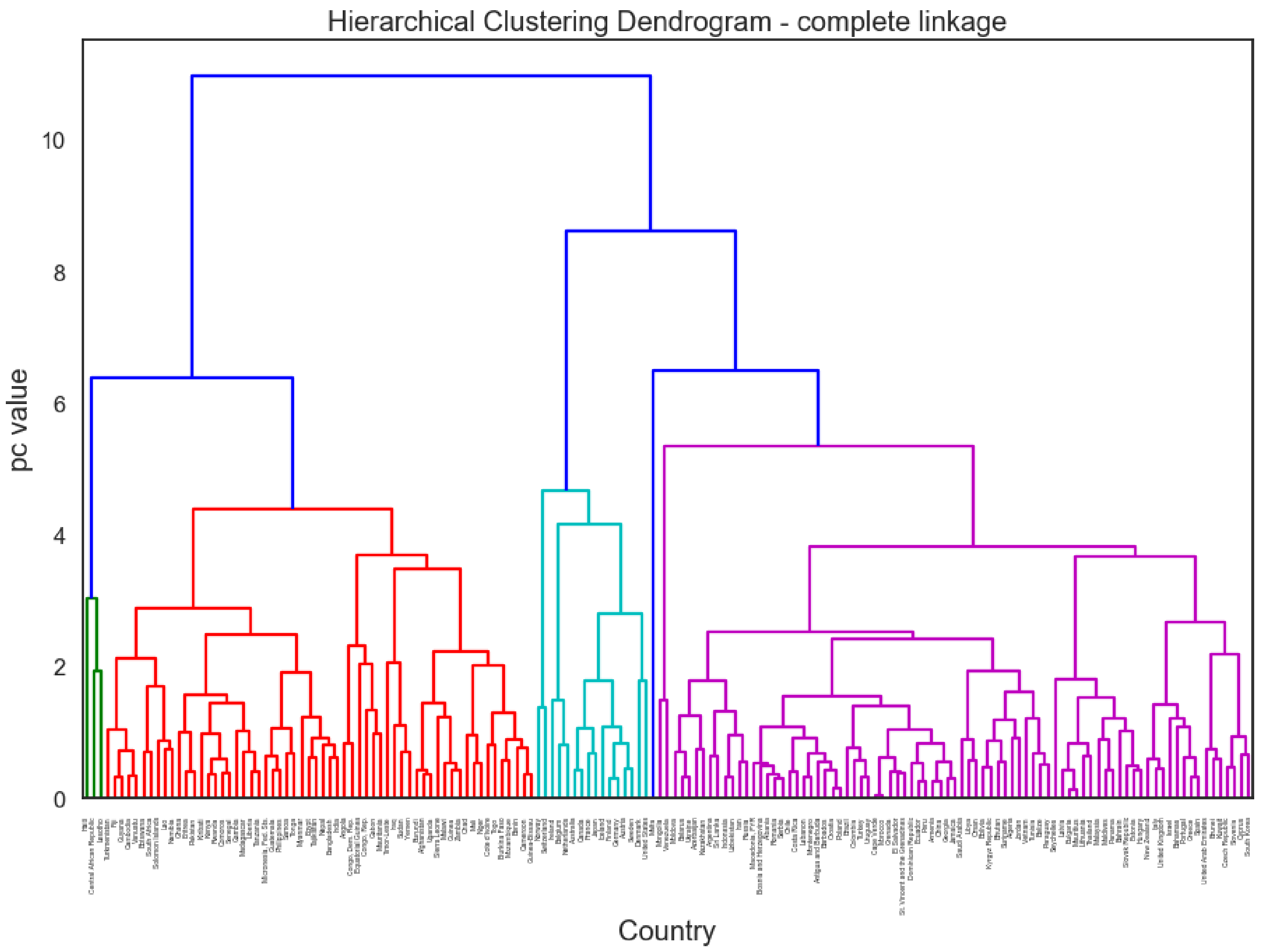
Hierarchical clustering (Linkage) - exclude outliers

As we can see from the graph of linkage dendrogram, it is not quite visible and doesn’t not suits properly with the dataset because we can cut the tree in a threshold value, we will use complete linkage dendrogram for hierarchical clustering.



Hierarchical clustering (Linkage) - exclude outliers

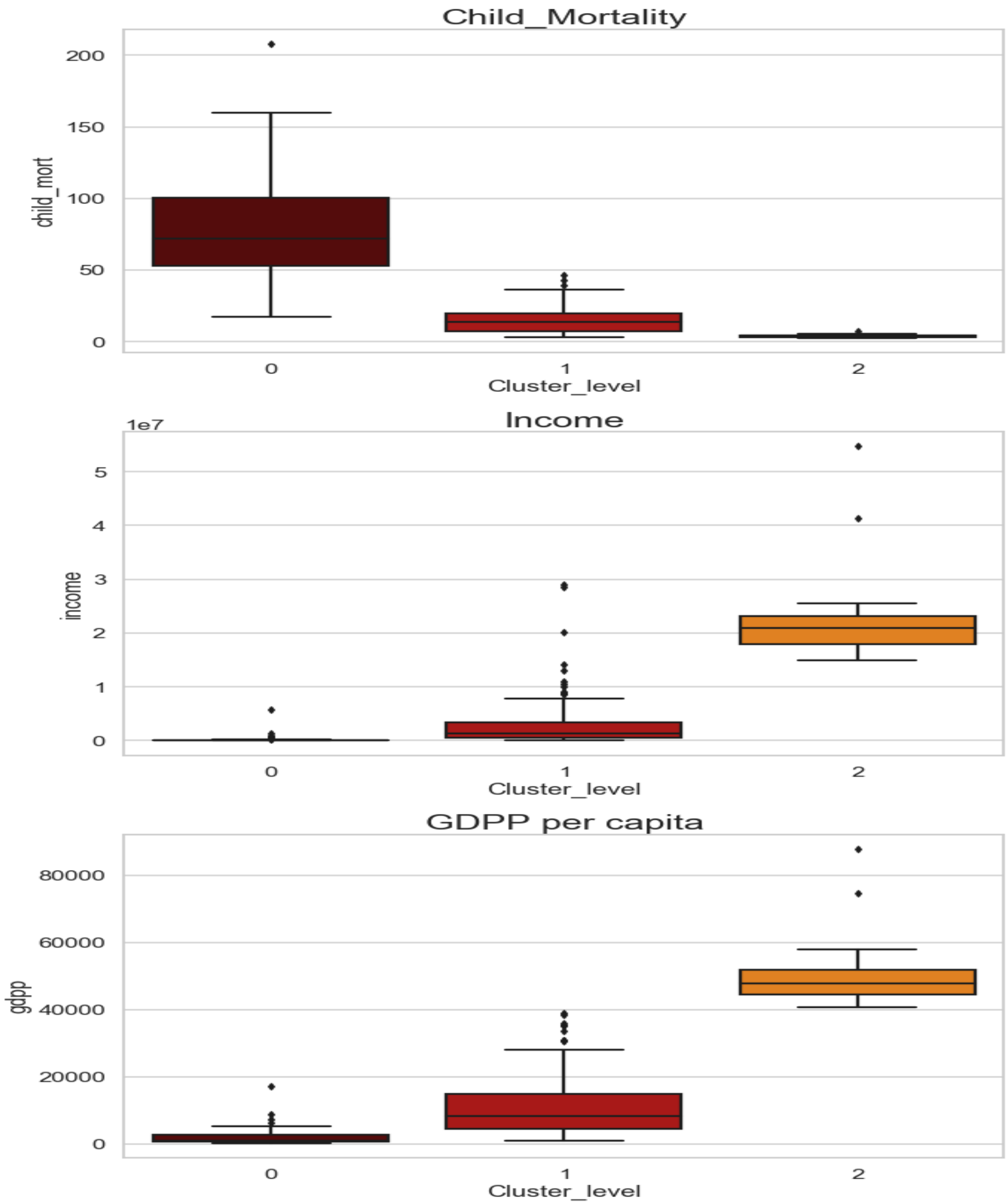
- Points noted from the graph on the right
- This graph shows proper way to decide number of clusters needs to be used by cutting at threshold value.
- We will cut at 3 branches which will give us 3 clusters



Visualization of original variables(Child mortality, Income and Gdpp)

Valuable Insights from three boxplots :

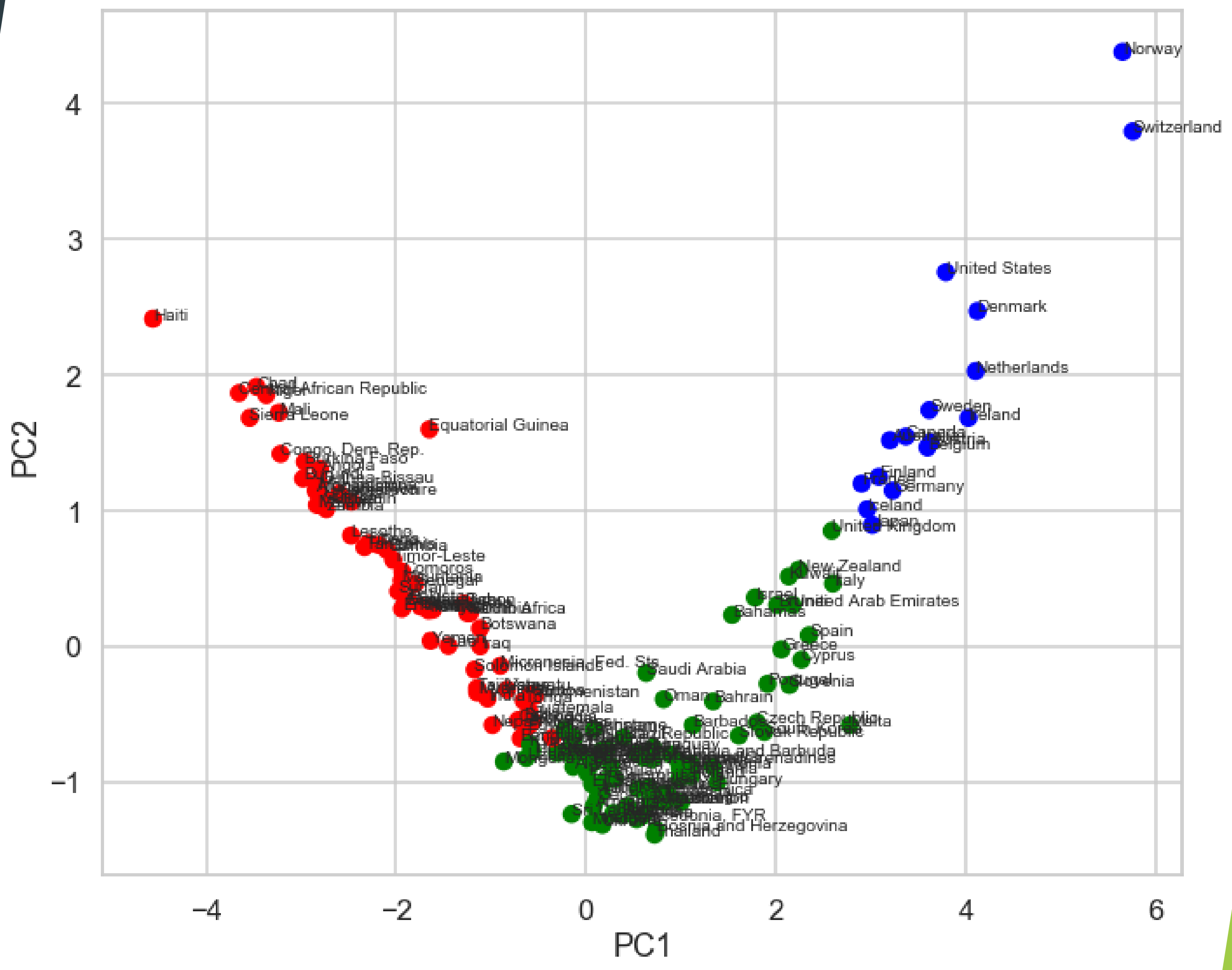
- - For cluster 0: .gdpp and income is the lowest than other clusters, Mortality of children is very high than other clusters.
- - For cluster 1: gdpp and income is having decent low value, mortality of children is high in here, the 4th quartile is larger than others
- - For cluster 2: gdpp and income is higher than other clusters, Mortality of children is very less compared to other clusters



Visualizing the PC1 and PC2 for hierarchical clustering - exclude outlier

As we can see from first two principal components(PC1 & PC2), the PC1 is in the direction where the countries need of least help. Here, why we are choosing PC1 because it has maximum percentage of variance explained.

The ‘Red’ color datapoints of countries need urgent help in aid but the 'Blue' one not required.



Insights - (Hierarchical : without outlier)

- Here, we got **63 countries** which need aid as they have having low income, high child mortality and low gdp per capita.
- Here, we got 16 countries which are having good social-economic and health factors.

Conclusion - Without outlier

K-Means vs Hierarchical Clustering

K-means clustering :

Countries that are direst need of aid

- Total 47 countries are in this category.

Hierarchical clustering :

Countries that are direst need of aid

- Total 63 countries are in this category.
- We have seen from both methods - (K-Means and Hierarchical clustering) that extra 9 countries are adding through hierarchical clustering. I would choose the final countries from hierarchical clustering as it gave **accurate** output than k-means clustering. I have compared the clusters and visualized from both methods and hierarchical clustering gave **precise** information than K-Means clustering.

Conclusion

- Among the two-conclusion drawn from approach 1 i.e. including outliers and approach 2 i.e. excluding outliers, approach 1 is the appropriate choice because it includes all the data points including outliers.
- As per the business requirements, we must find all the countries which are in direst need of aid i.e. the countries which are having low socio-economic and health factors. Hence, we can't exclude any countries from our dataset as it will create a major drawback in our model.
- For example, let's take an outlier country 'Nigeria' which is having low socio-economic and health factors. If we exclude this outlier from my dataset, we will miss our main objective as it happened with approach 2. So, even though the model was greater than the previous model, we can't use it as it doesn't suit the business needs.
- Selecting approach 2 means we must loose many countries in process which is not ideal from business perspective.
- The final list of 47 countries name needs to focus on the most are mentioned below :
- Afghanistan, Angola, Benin, Botswana, Burkina Faso, Burundi, Cameroon, Central African Republic, Chad, Comoros, Congo, Dem. Rep., Congo, Rep., Cote d'Ivoire, Equatorial Guinea, Eritrea, Gabon, Gambia, Ghana, Guinea, Guinea-Bissau, Haiti, Iraq, Kenya, Kiribati, Lao, Lesotho, Liberia, Madagascar, Malawi, Mali, Mauritania, Mozambique, Namibia, Niger, Nigeria, Pakistan, Rwanda, Senegal, Sierra Leone, South Africa, Sudan, Tanzania, Timor-Leste, Togo, Uganda, Yemen and Zambia

Thank You