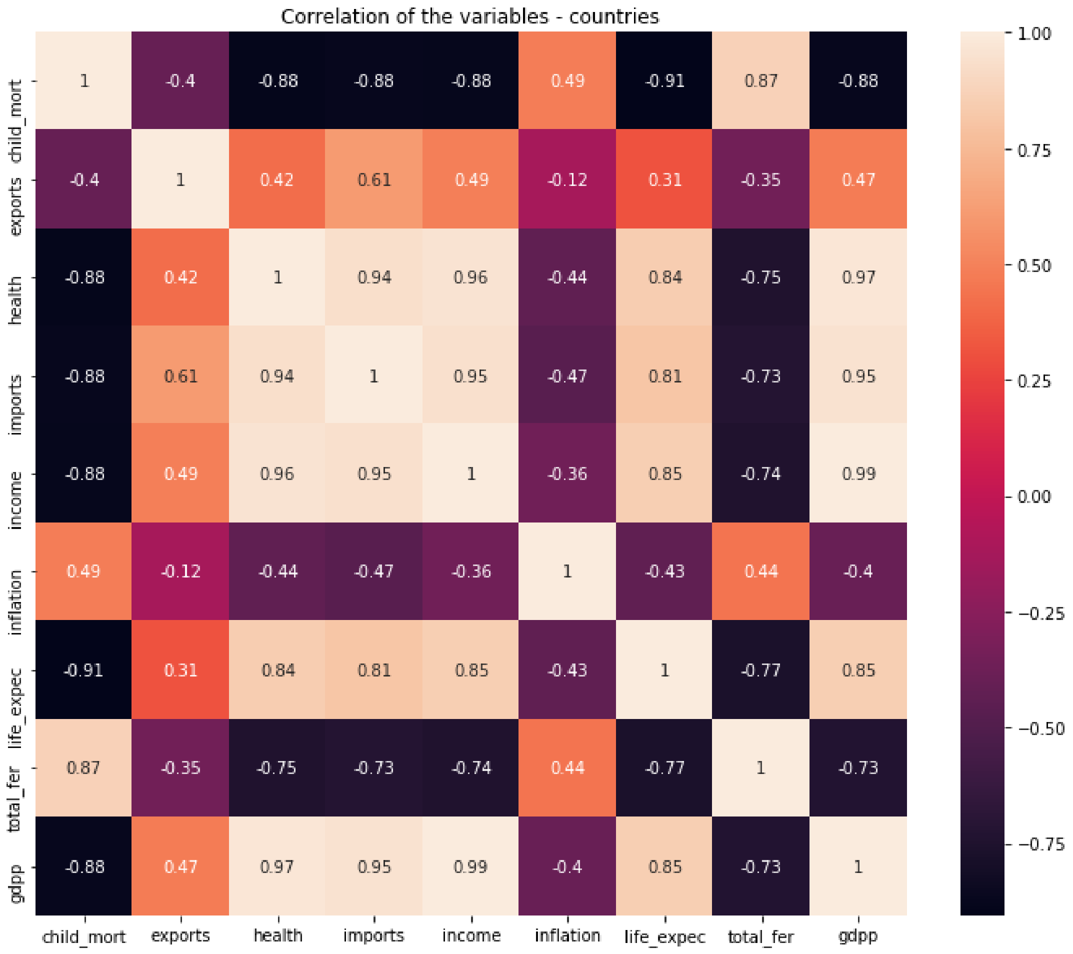
Clustering of Countries

CATEGORISE THE COUNTRIES USING SOME SOCIO-ECONOMIC AND HEALTH FACTORS THAT DETERMINE THE OVERALL DEVELOPMENT OF THE COUNTRY

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.



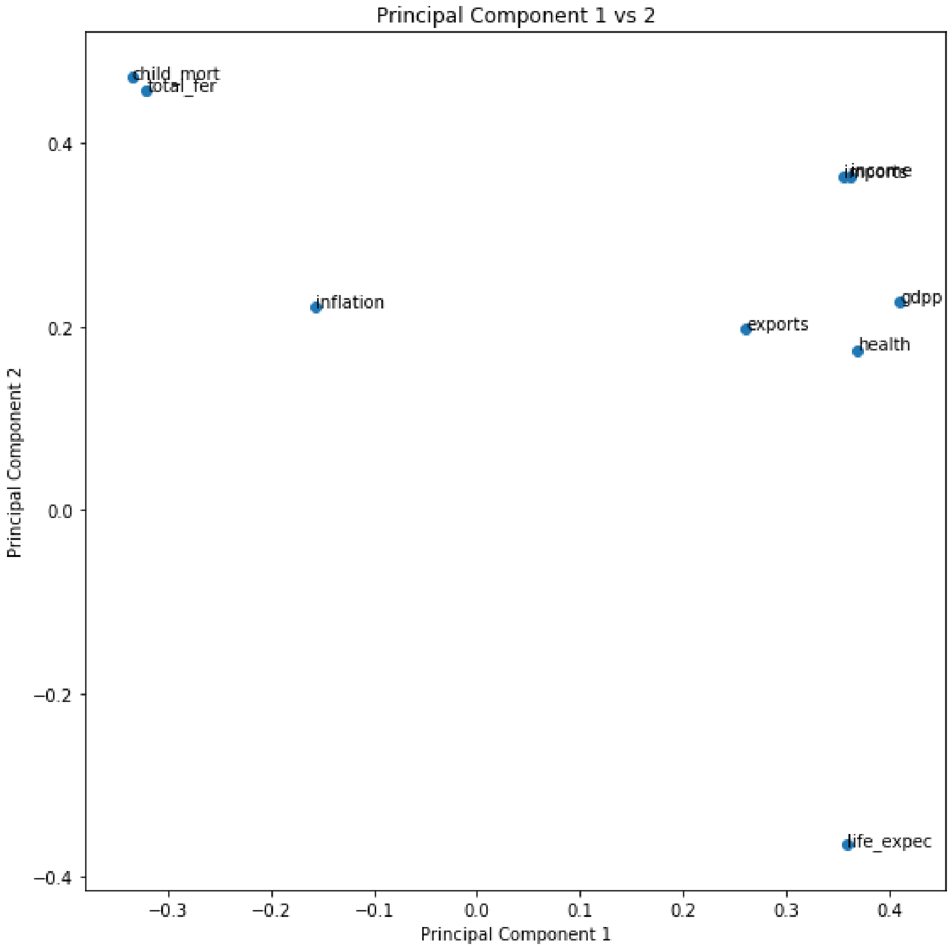
Correlation of the variables - countries

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

PCA Components(PC1 vs PC2)

From the plot on the left we can conclude the below points :-

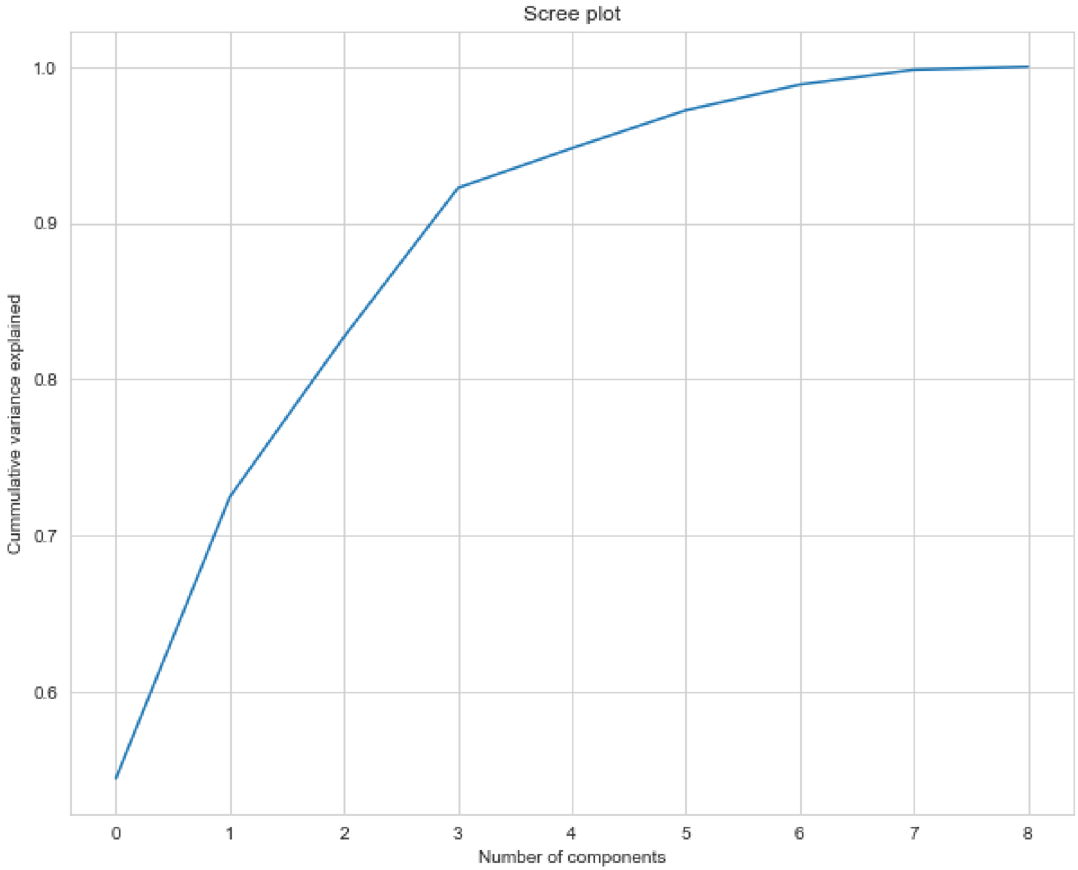
- The data point of life expectancy is in the direction of principal component 1.
- All the high value data points are in the direction of principal component 1.



PCA – Scree plot

From the graph on the left points to be concluded –

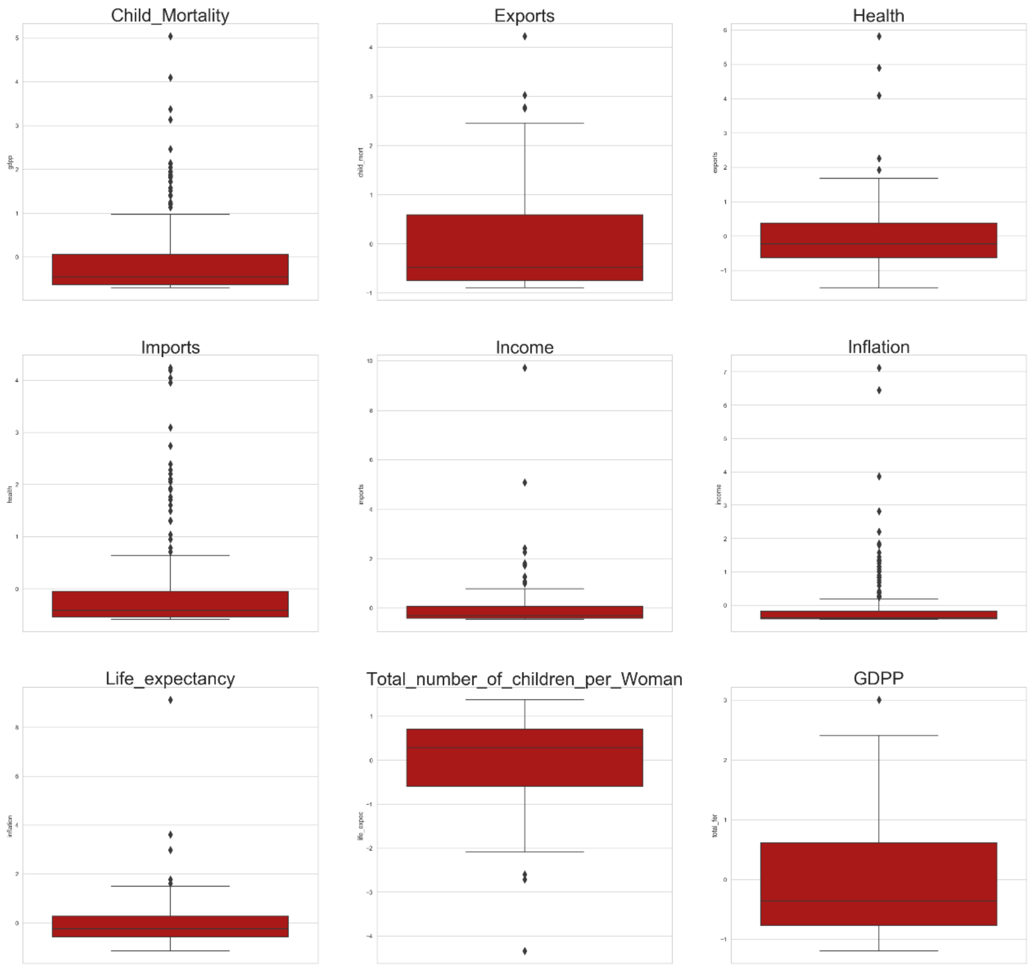
- The number of components equal to 4 is having approx. 95% of variance explained.
- The number of component equal to 5 is having approx. 97% of variance explained.
- So, the ideal number of components can be chosen is 5.



Visualization of outliers

From the boxplot attached to the left, points to be concluded –

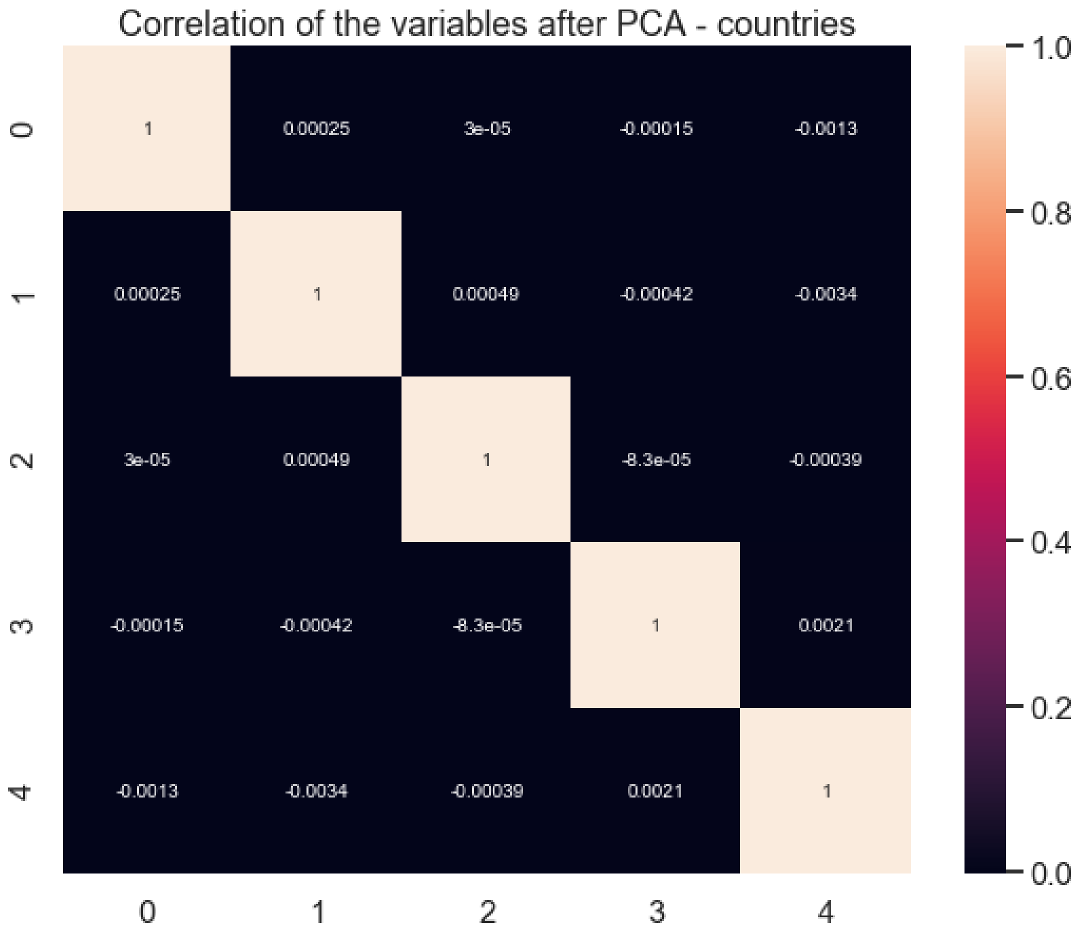
- As we can see all the boxplot created for the variables are having different amount of outliers.
- All boxplots except one ‘Total_number_of_children_per_women’ is having outliers on the bottom of the boxplots which means there are some countries where no. of children per women is very less compared 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 colour, which means they all are close to 0.
- This show 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

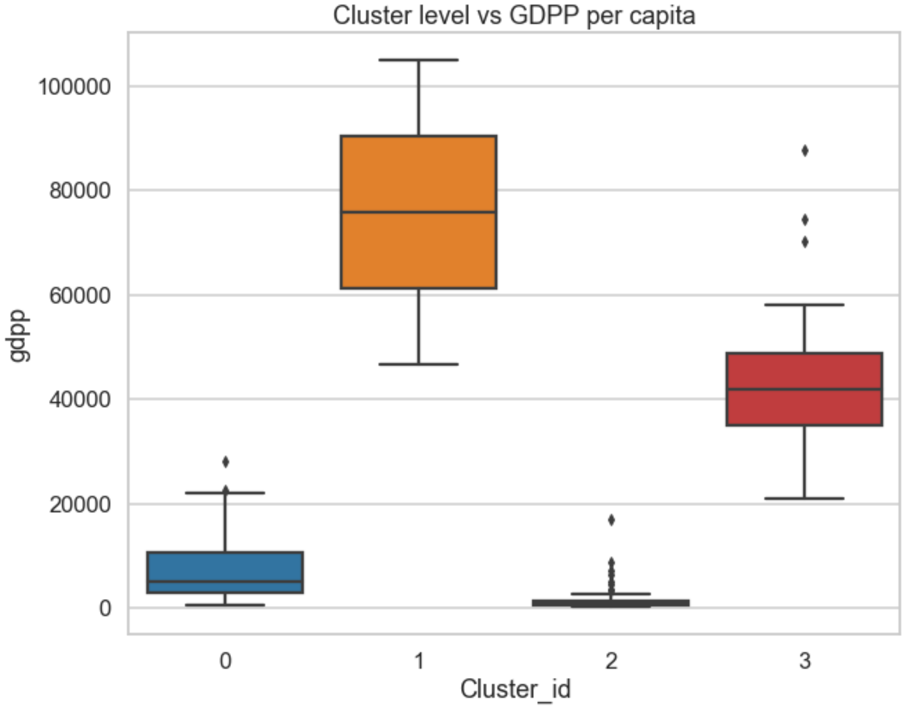
K-means analysis

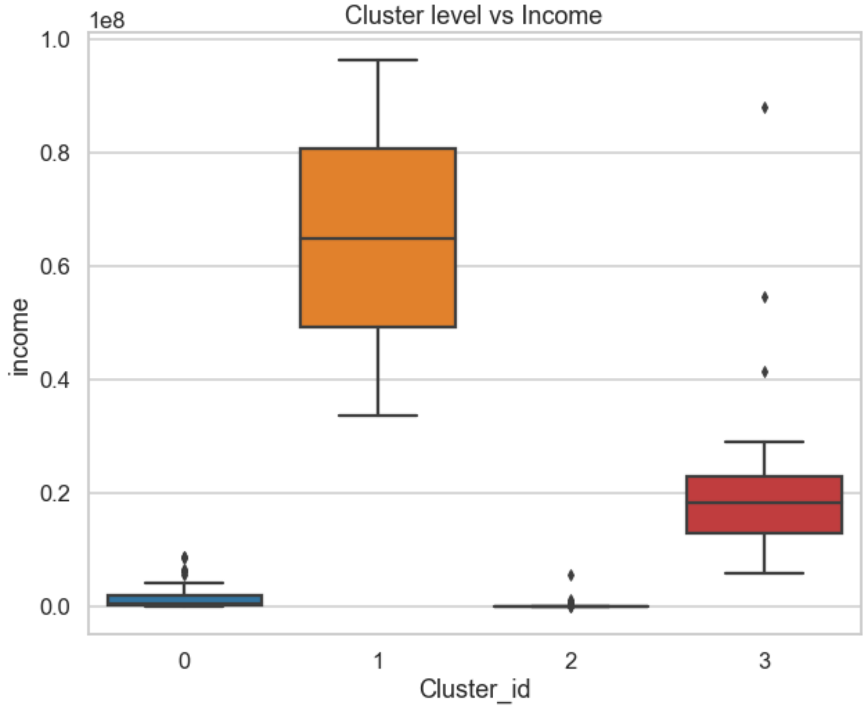
- We are trying to find the optimum value of the k-mean based on the business requirements.
- So to achieve this we used silhouette analysis to find the score of range of cluster values.
- We found the below silhouette scores:

```
 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 0.2970430701052886
For no. of cluster=6,silhouette score is 0.3885620385271955
For no. of cluster=7,silhouette score is 0.2790924078029563
For no. of cluster=8,silhouette score is 0.2869297133457836
For no. of cluster=9,silhouette score is 0.24637804990944706.
```

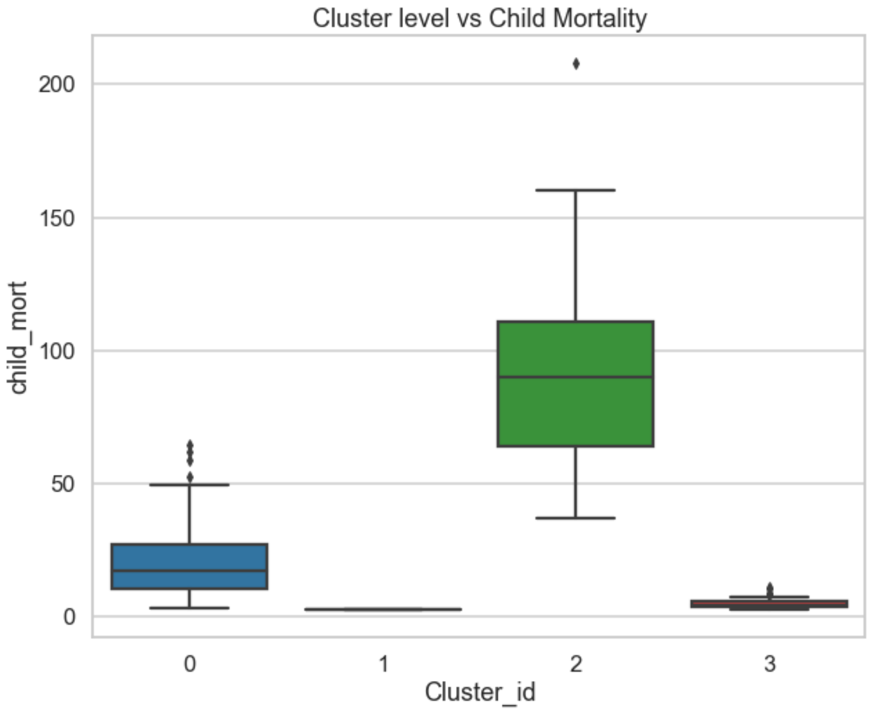
- We found that cluster = 2 is having highest score hence k value should be 2 but, if we choose k value as ‘2’, it will not suit business needs.
- Hence we use k value as ‘4’ since it is giving precise information also fulfils business needs.

Visualization of the original variable with clusters





Visualization of the original variable with clusters



Valuable Insights from boxplot shown on the previous two slides – gdpp, income and child mortality

Points to be concluded from above boxplot - gdpp(The GDP per capita) visualization

- Cluster label 0 : Most of the countries in this boxplot are having little high gdpp than the very lowest gdpp group
- Cluster label 1 : Having highest gdpp than all other cluster labels
- Cluster label 2 : Having very low gdpp than all the other cluster labels
- Cluster label 3 : Having high gdpp with some outliers

Valuable Insights from boxplot shown on the previous two slides – gdpp, income and child mortality

Points to be concluded from above boxplot - income(Net income per person) visualization

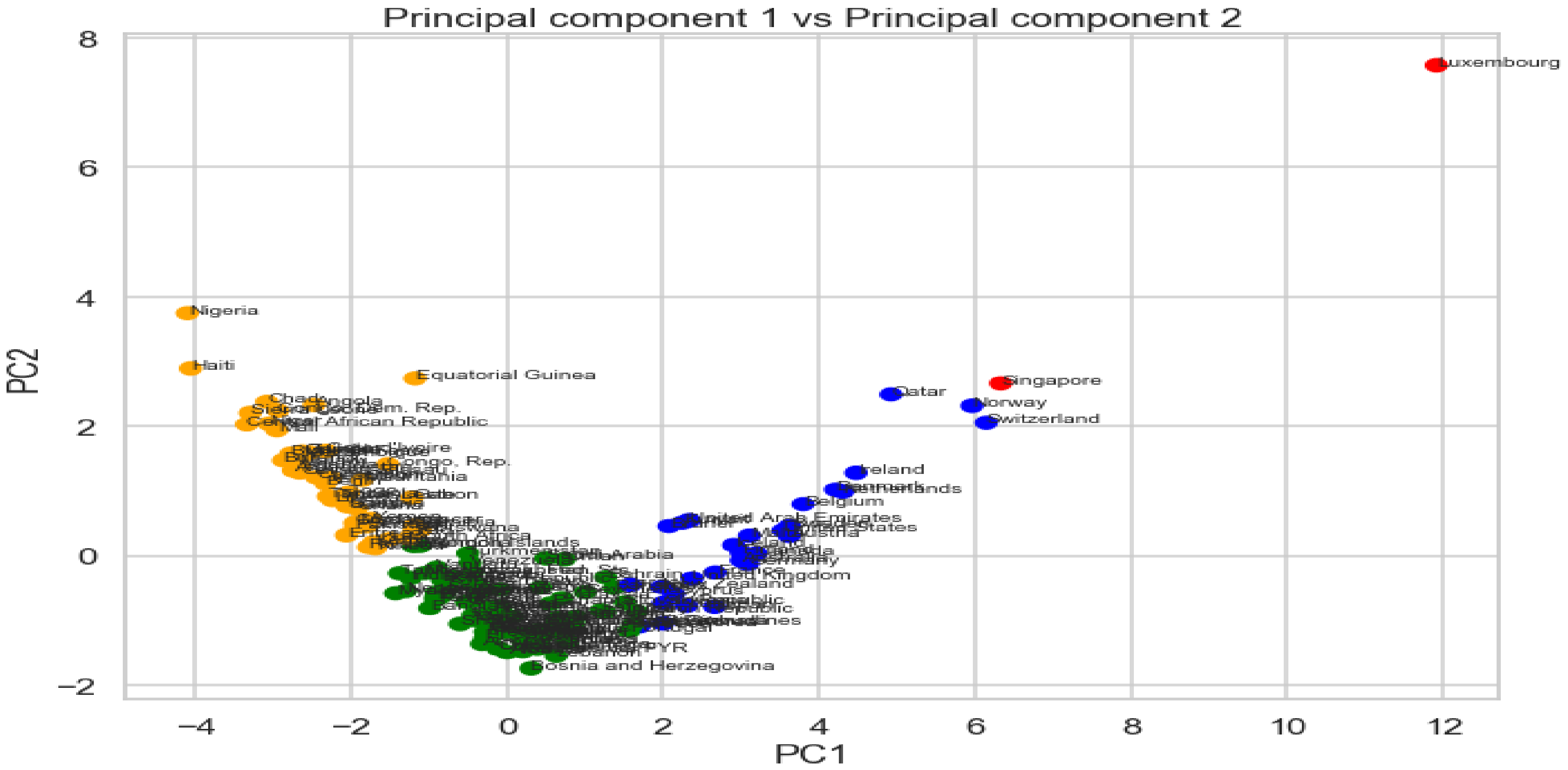
- Cluster label 0 : Having second lowest income with few outliers
- Cluster label 1 : Having highest income than all other cluster labels
- Cluster label 2 : Having lowest income with few outliers than rest of the cluster labels
- Cluster label 3 : Having decent income with few outliers

Valuable Insights from boxplot shown on the previous two slides – gdpp, income and child mortality

Points to be concluded from above boxplot - child_mort(Death of children under 5 years of age per 1000 live births) visualization

- Cluster label 0 : Having second highest child mortality with few outliers
- Cluster label 1 : Having lowest child mortality than all other cluster labels
- Cluster label 2 : Having highest child mortality than rest of the cluster labels
- Cluster label 3 : Having low child mortality with few outliers

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



Insights drawn from the scatter plot visualization

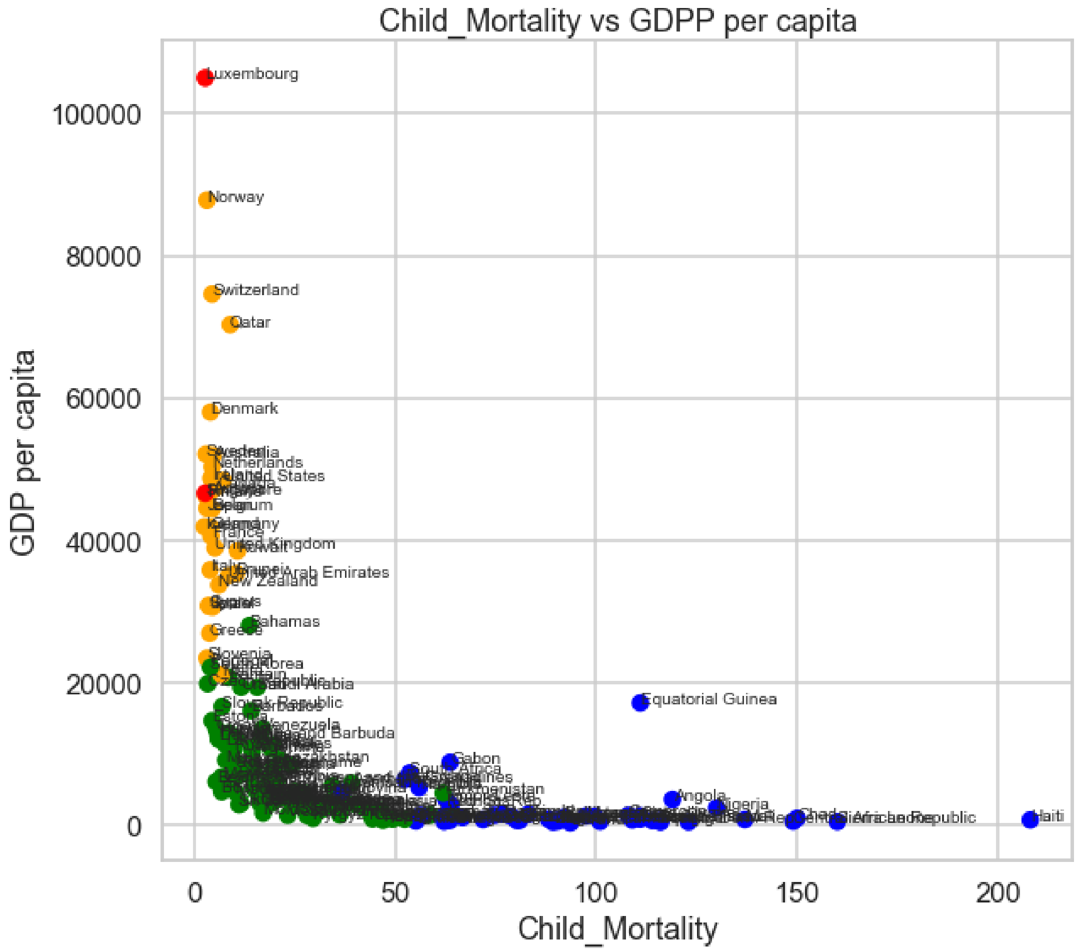
We can conclude that:-

- 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 'Qatar' and 'Norway' are having high PC1 which means they are doing well
- Where on other hand countries like 'Singapore‘ and ‘Luxembourg’ requires urgent need of aid.

Visualizing with original variables (Child_mort vs gdpp)

From the above scatter plot, Two points we can conclude :

- Country "Singapore" and " Luxembourg" is in dire need of aid
- Country "Norway" is having good gdp and less child mortality rate



Insights (K-Means with outliers)

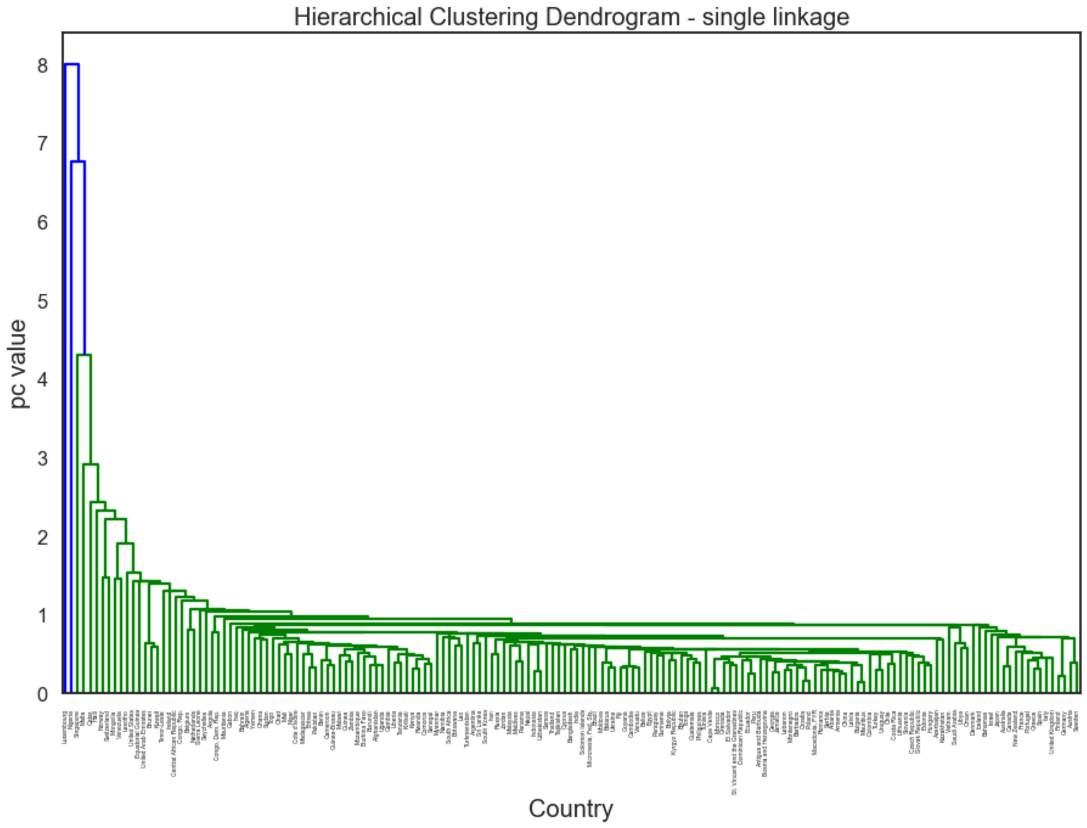
- There are total 2 countries from the dataset need of urgent help/aid as they are having lowest income, high child mortality and low gdp per capita.

- 29 countries is there with good socio-economic and health factors. These are countries with highest income, low child mortality and high gdp per capita or the countries that doesn't need aid.

Hierarchical clustering(Linkage)

This is another method to find our low development countries.

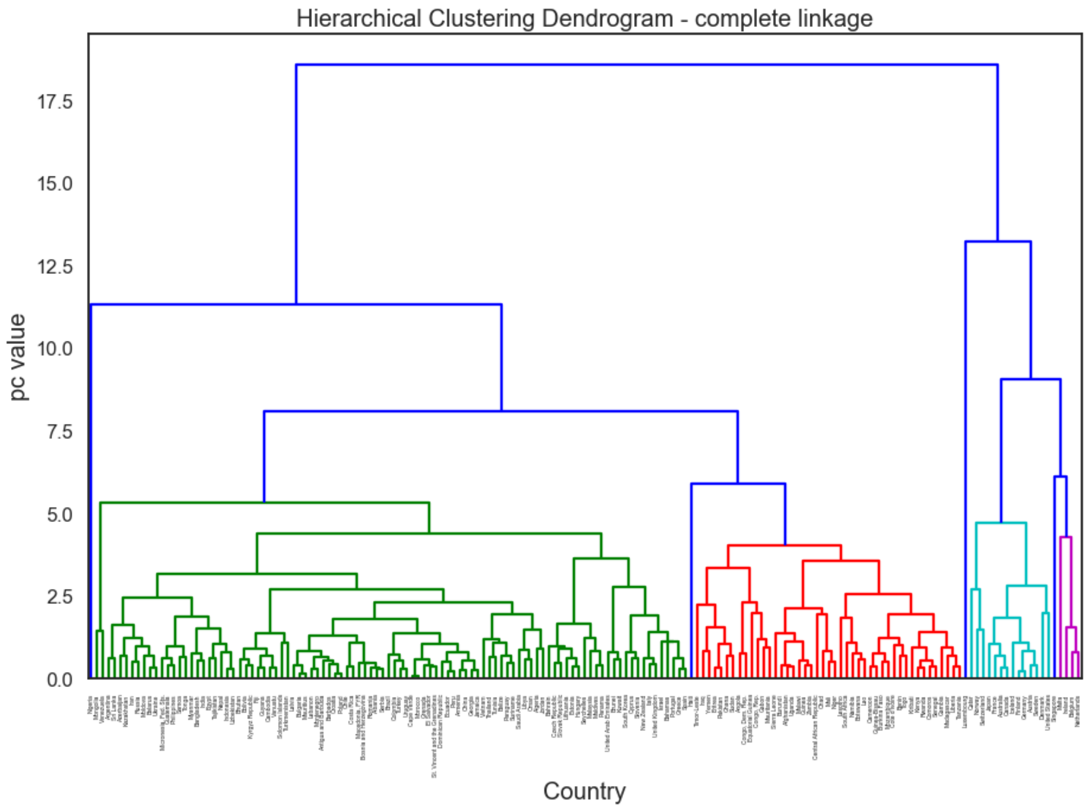
As we can see from the graph of linkage dendrogram, it is not quite visible and doesn’t not suits properly with the dataset because we can cut the tree in a threshold value, we will use complete linkage dendrogram for hierarchial clustering.



Hierarchical clustering(complete linkage)

Points noted from the graph on the right –

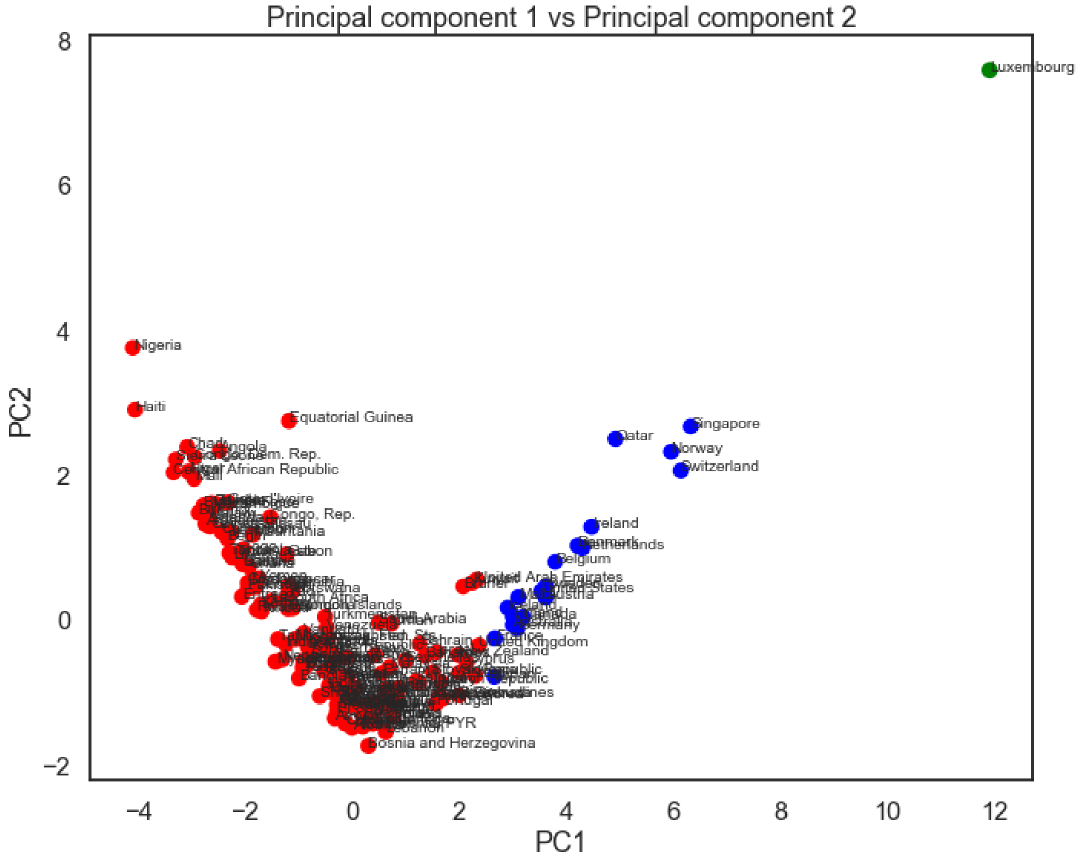
- This graph shows proper way to decide number of clusters needs to be used by cutting at threshold value.
- We will cut at 3 branches which will give us 3 clusters.



Visualization of hierarchical clustering with first two principal components

Points to be drawn from the hierarchical scatter plot –

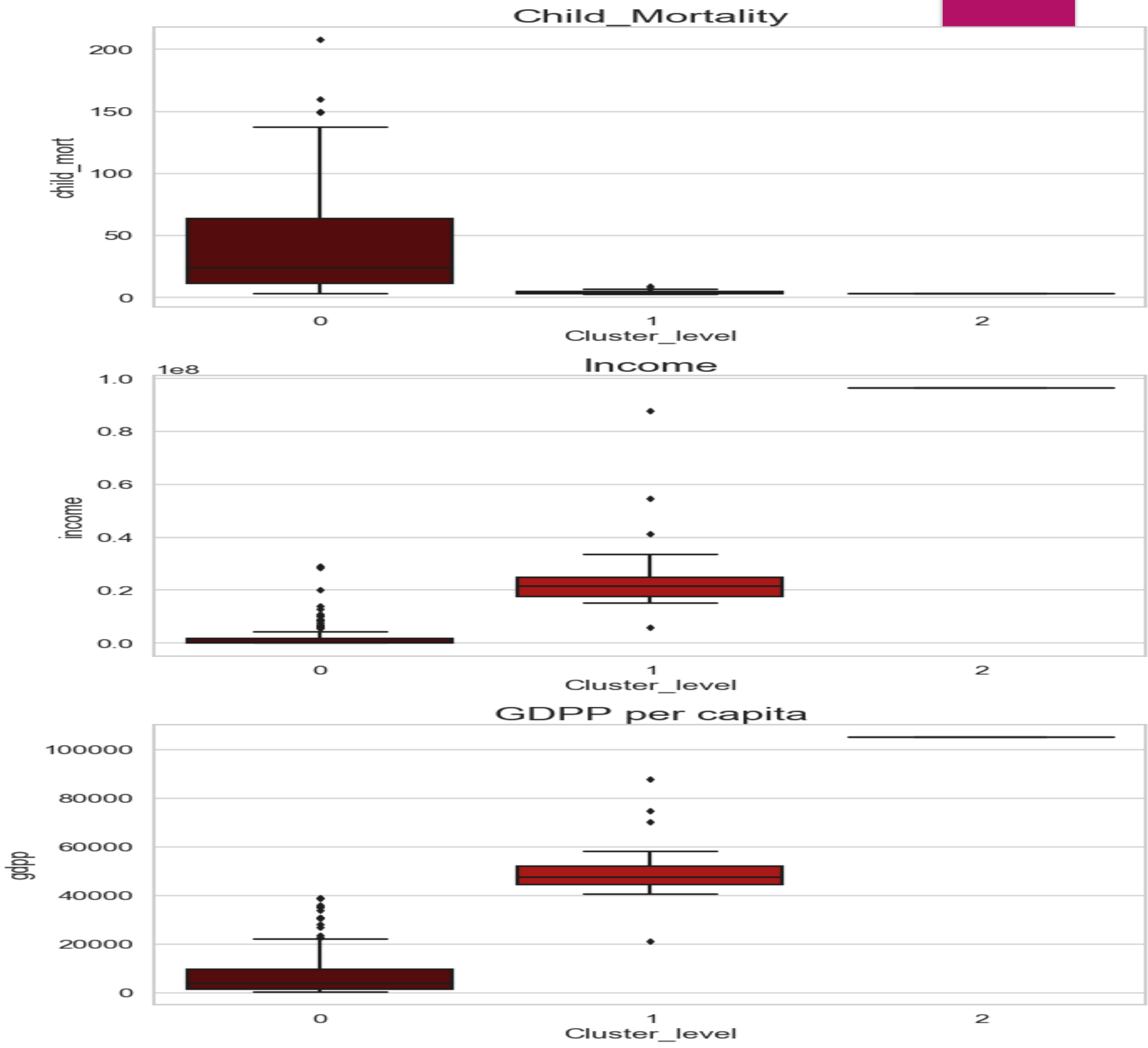
- As we can see from first two principal components(PC1 & PC2), the PC1 is in the direction where the countries need of least help. Here, why we are choosing PC1 because it has maximum percentage of variance explained.
- The 'Red' color datapoints of countries need help in aid but the 'Blue' one not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



Visualization of original variables(Child mort, 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)

Countries that are direst need of aid -Total 147 countries are in this category, those are in need of urgent help/aid it is having lowest income, high child mortality and low gdp per capita

Countries that are having good socio-economic and health factors -1 country is in this category - Luxembourg

Conclusion – With outliers

K-Means vs Hierarchical Clustering

K-means clustering :

- Countries that are direst need of aid -Total 2 countries are in this category - Luxembourg and Singapore
- Countries that are having good socio-economic and health factors -Total 29 countries are in this category

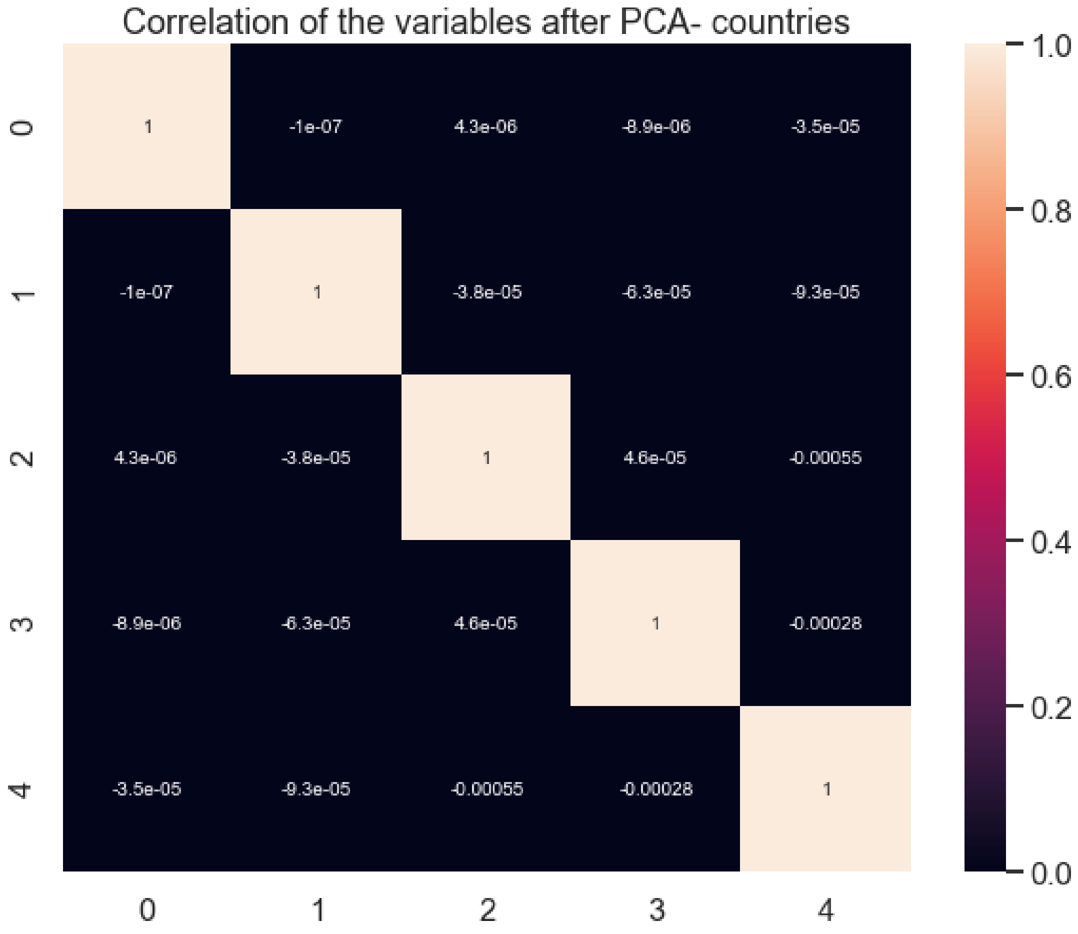
Hierarchical clustering :

- 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

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.

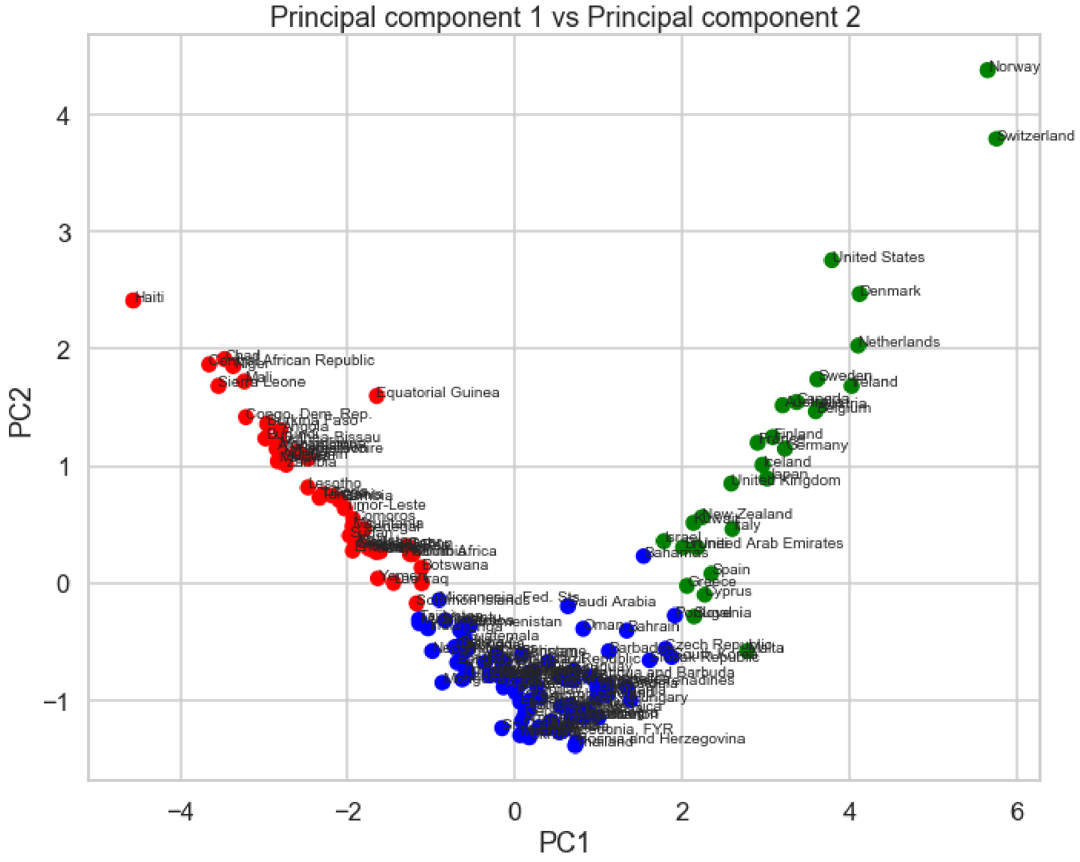


Correlation of the variables after PCA- countries

	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 chosing 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.



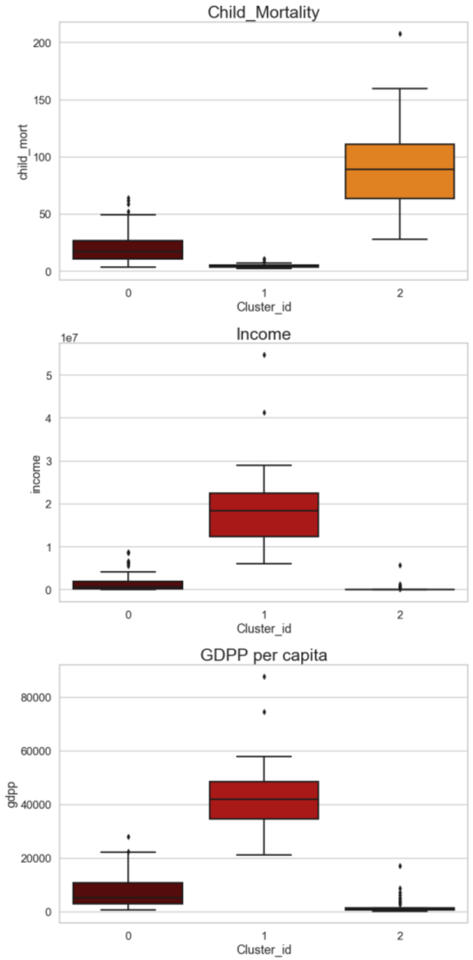
Visualisation of original variables(gdpp, income and child_mort)

Valuable Insights from above 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.



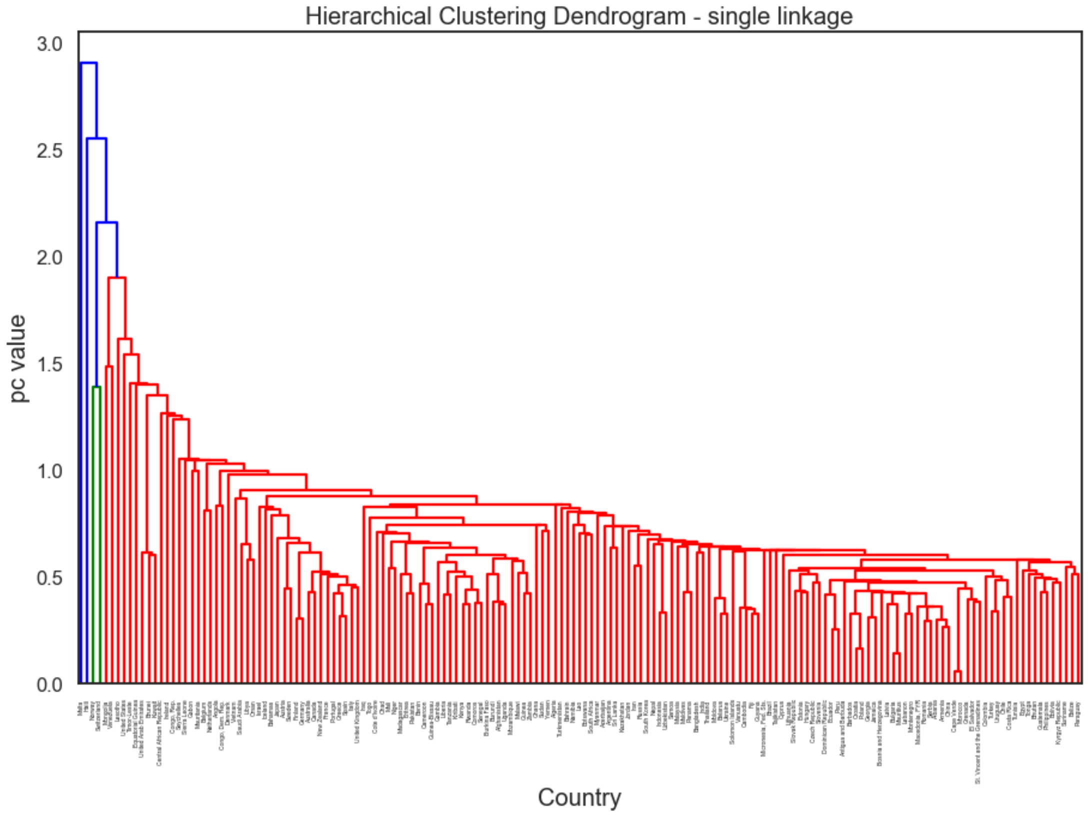
Insights (K- Means exclude outliers)

- There are total 47 countries from the dataset are in 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.

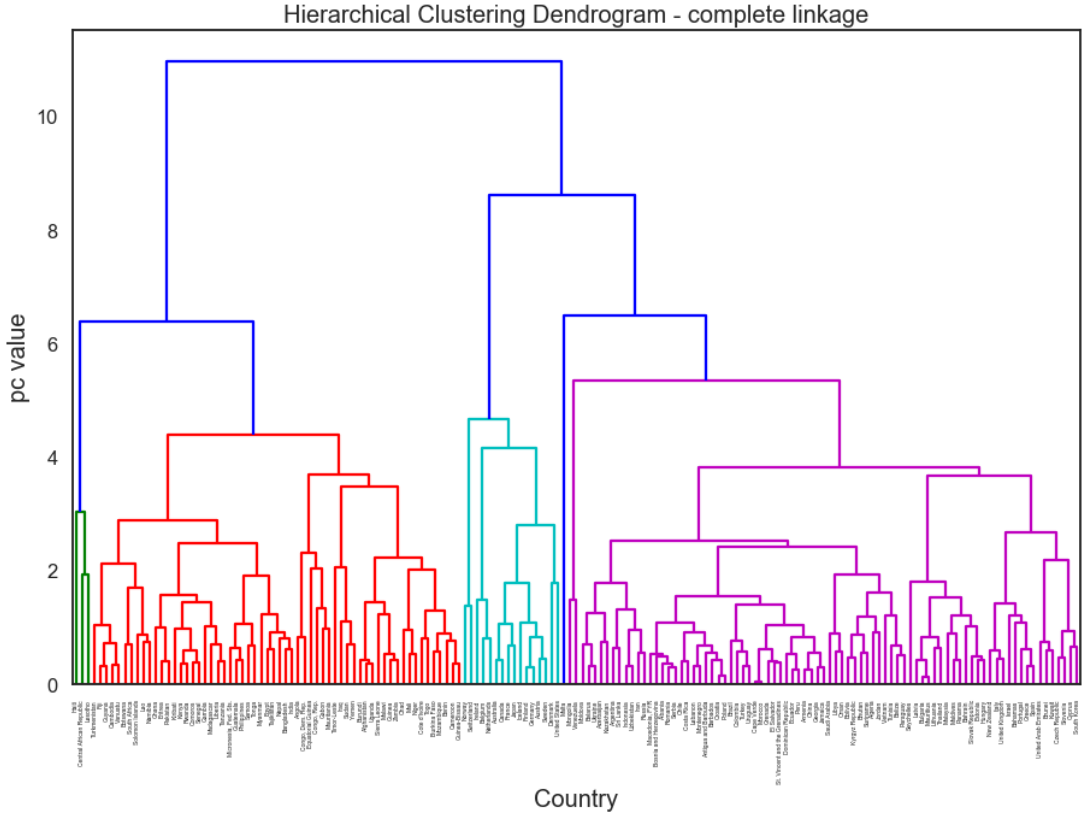
Hierarchical clustering(Linkage) – exclude outliers

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



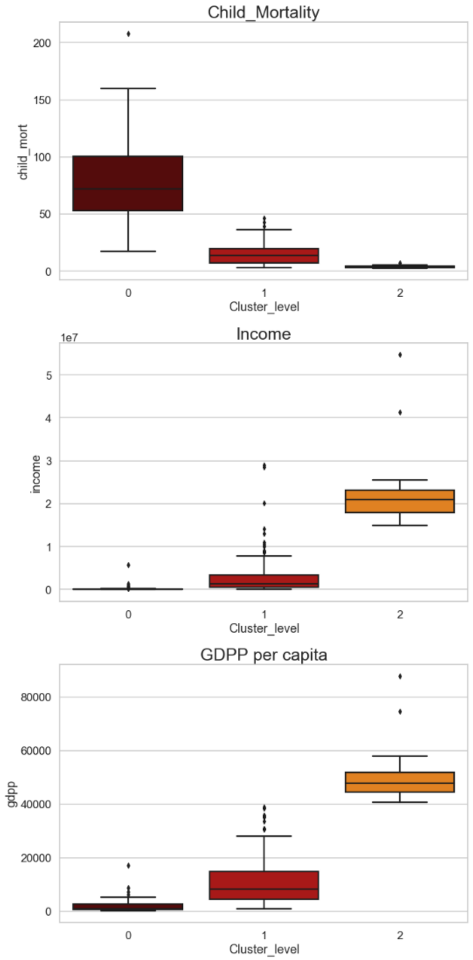
Visualization of original variables(Child mortality, Income and Gdpp)

Valuable Insights from above three boxplots :

For cluster 0: .gdpp and income is the lowest than others clusters, Mortality of children is very high than other clusters.

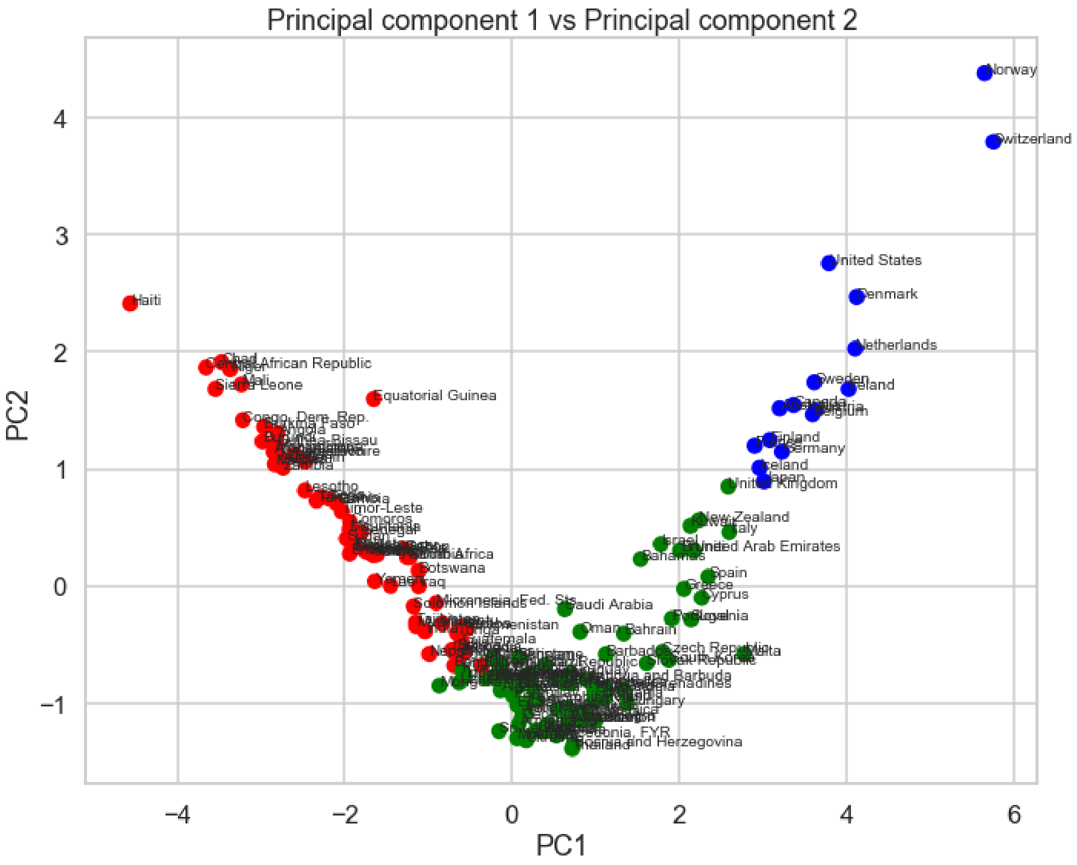
For cluster 1: gdpp and income is having decent low value, mortality of children is high in here, the 4th quartile is larger than others

For cluster 2: gdpp and income is higher than others 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 Approach) exclude outliers

- We got 63 countries which are in need of aid as they have having low income, high child mortality and low gdp per capita.

- We got 16 countries which are having good social-economic and health factors.

Conclusion – exclude outliers

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.

Final Conclusion

- Among the two conclusion drawn from approach 1 i.e. including ouliers 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 have to 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. 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 suits the business needs.
- Selecting approach 2 means we have to loose many countries in process which is not ideal from business prespective.
- The final list of 2 countries name needs to focus on the most are mentioned below :

Singapore

Luxembourg