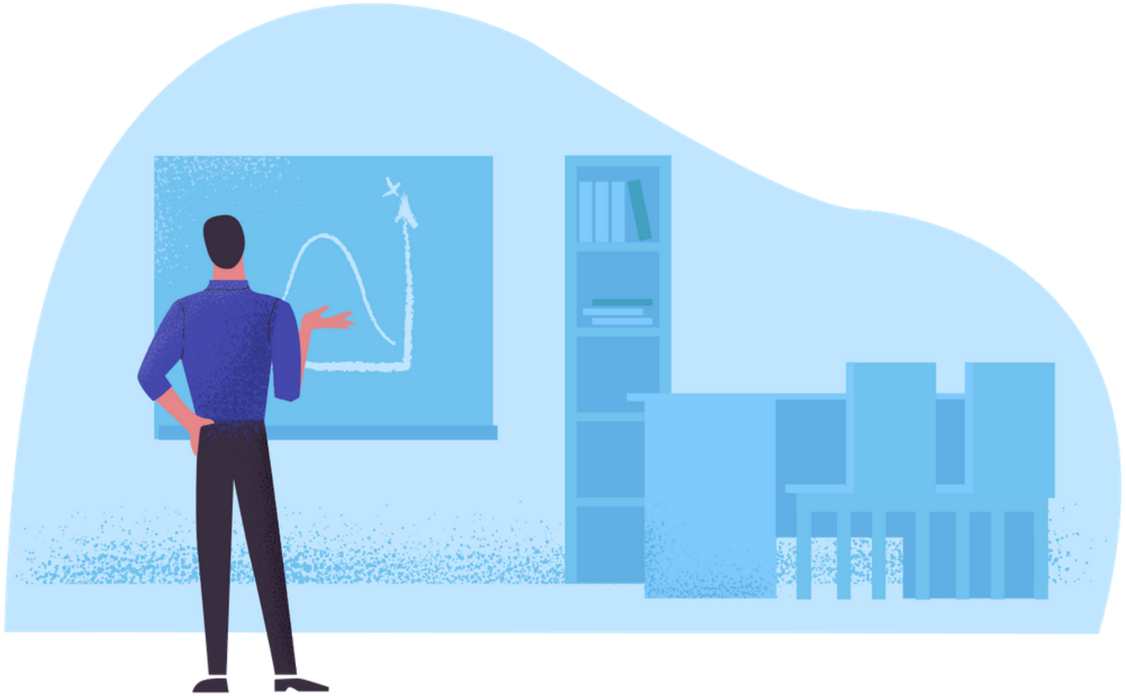
CLUSTERING ASSIGNMENT



BY

SHUBHNEET ARORA

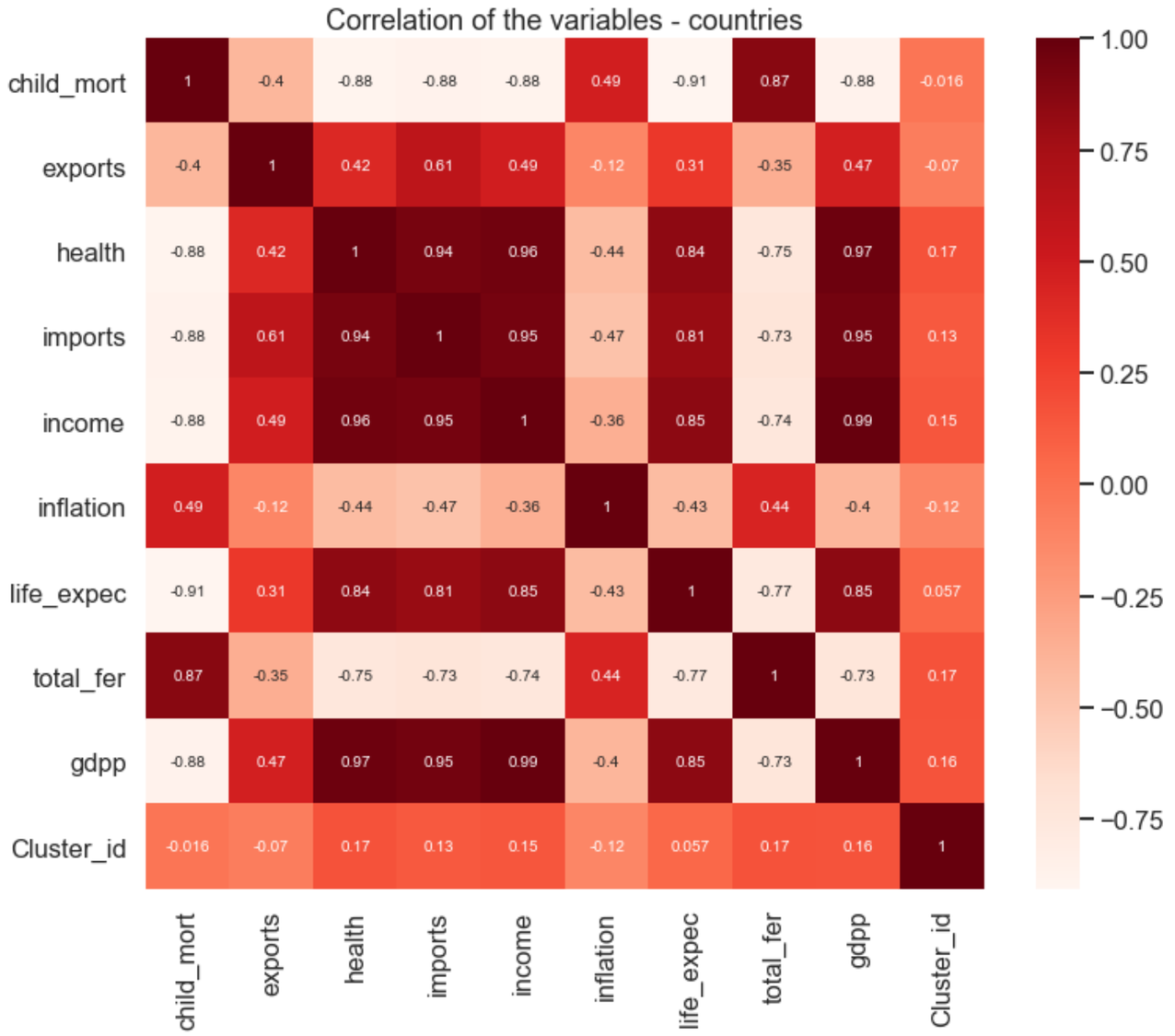
Correlation of variables

INFERENCES:

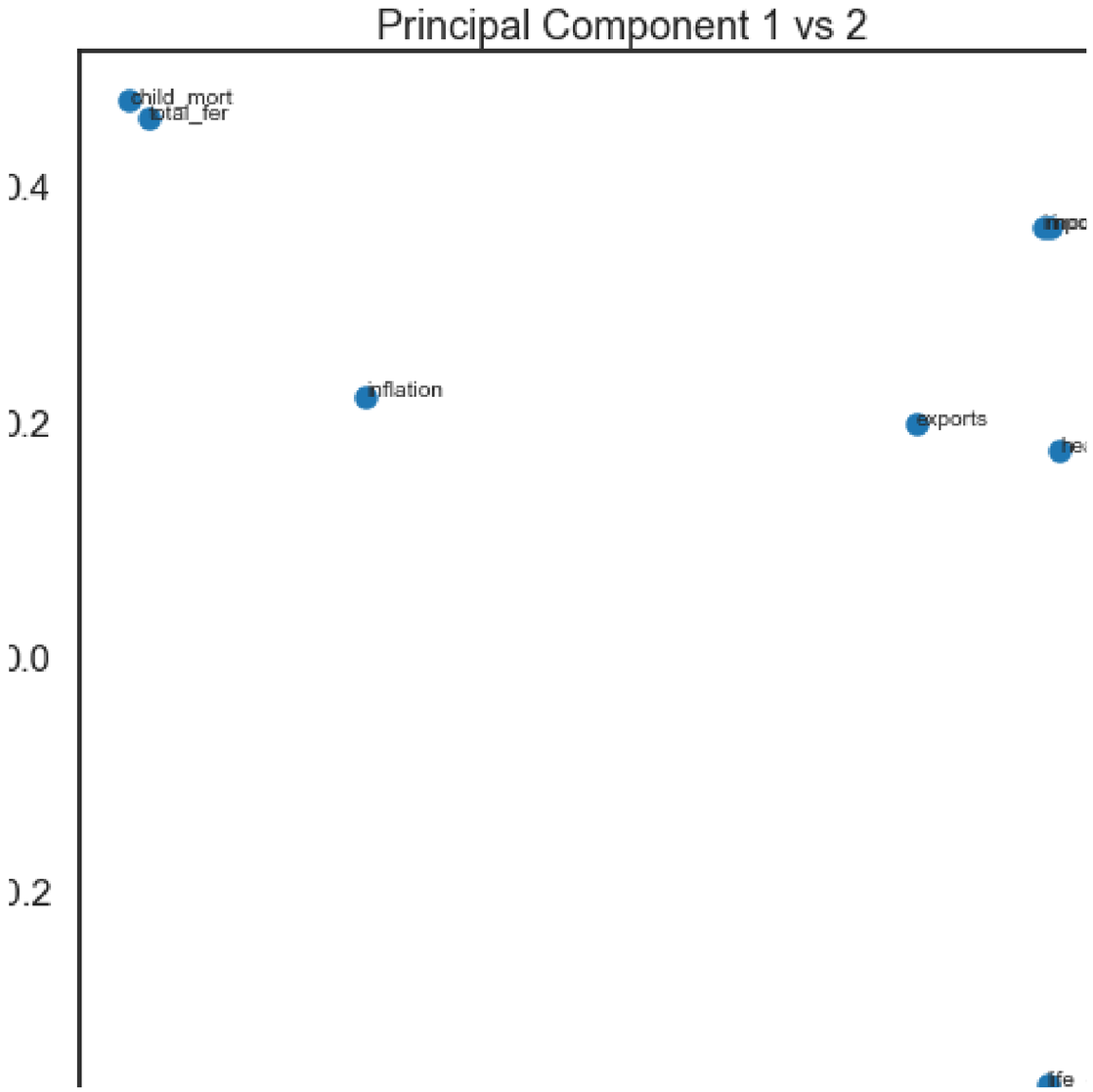
1.some of the variables
have a high positive correlation like
income, health, imports.

2.some of the variables
have a high negative correlation like
child mortality and life expectancy,
gdpp, health
etc.

3.this shows that the data set
is having multicollinearity.



PCA components

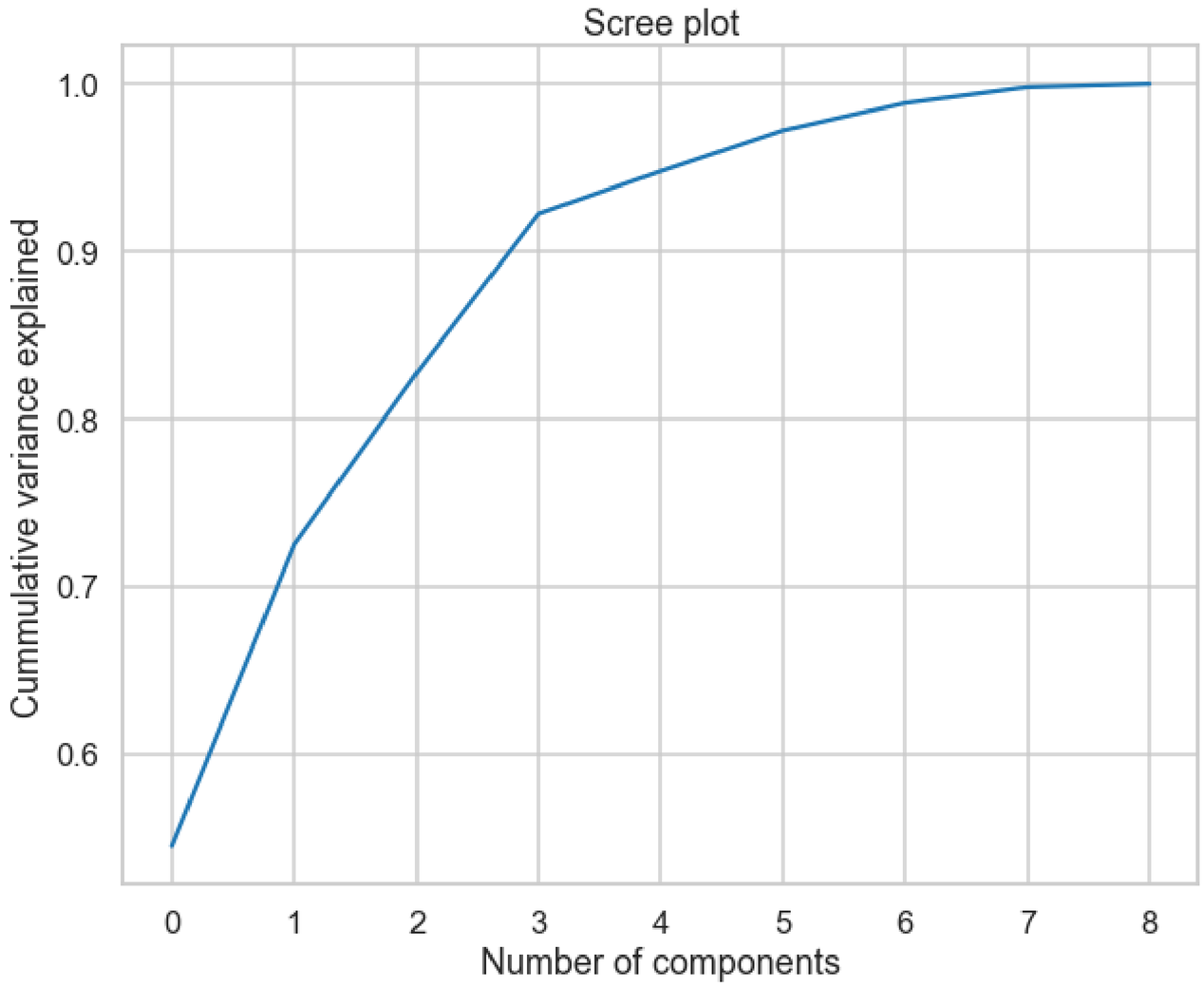


INFERENCES

The data points of life expectancy are in the direction of principal component 1.

The data points of child mortality and total fertility is in the direction of principal component 2.

PCA - Scree plot



INFERENCES

The number of component equal to 4 is having approx. 95% of variance explained.

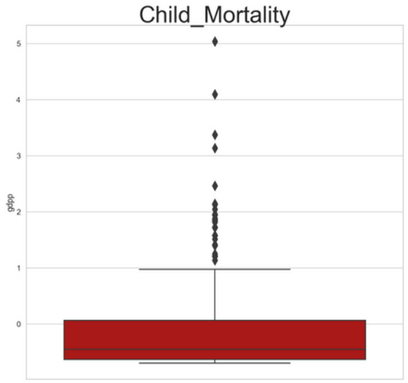
The number of component equal to 5 is having approx. 97% of variance explained.

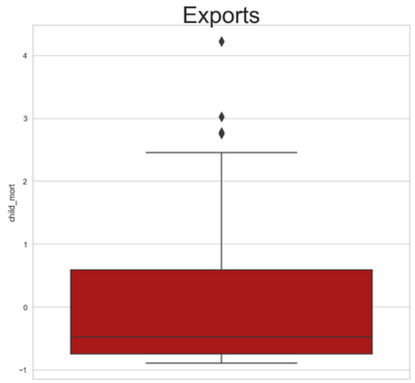
So, the ideal number of components can be chosen is 5

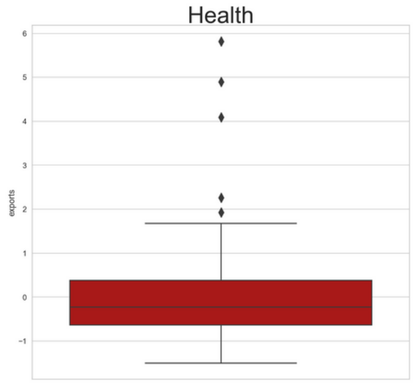
Visualization of Outliers

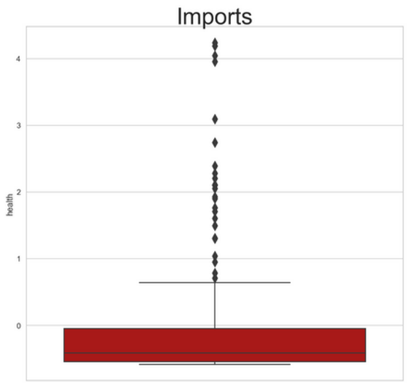
INFERENCES

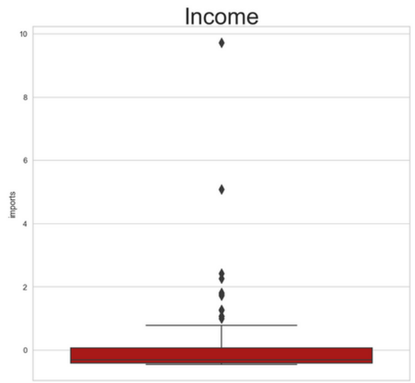
As we can see that all
the boxplots created for the
variables are having decent
number of outliers.
All boxplots except one
'Total_number_of_children_per_
woman'
is having outliers on the bottom of
the boxplots which means there
are some countries
were no. of children per woman is
very less compared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

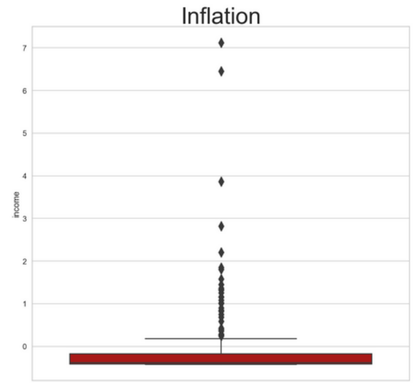


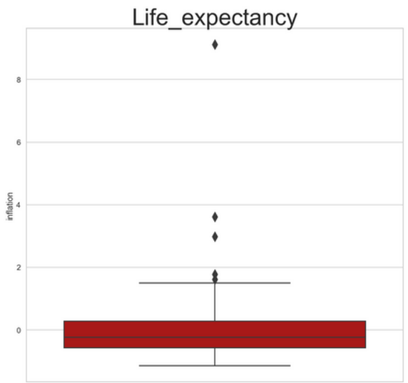


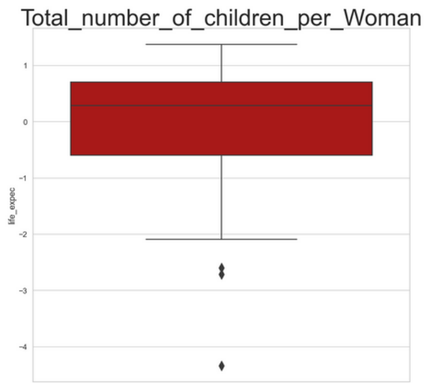


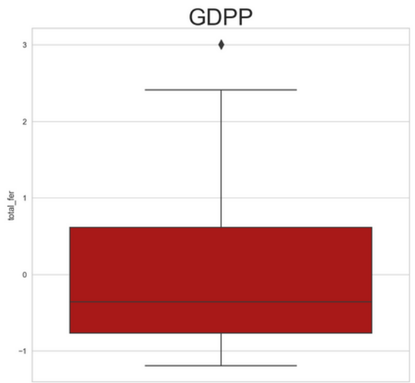








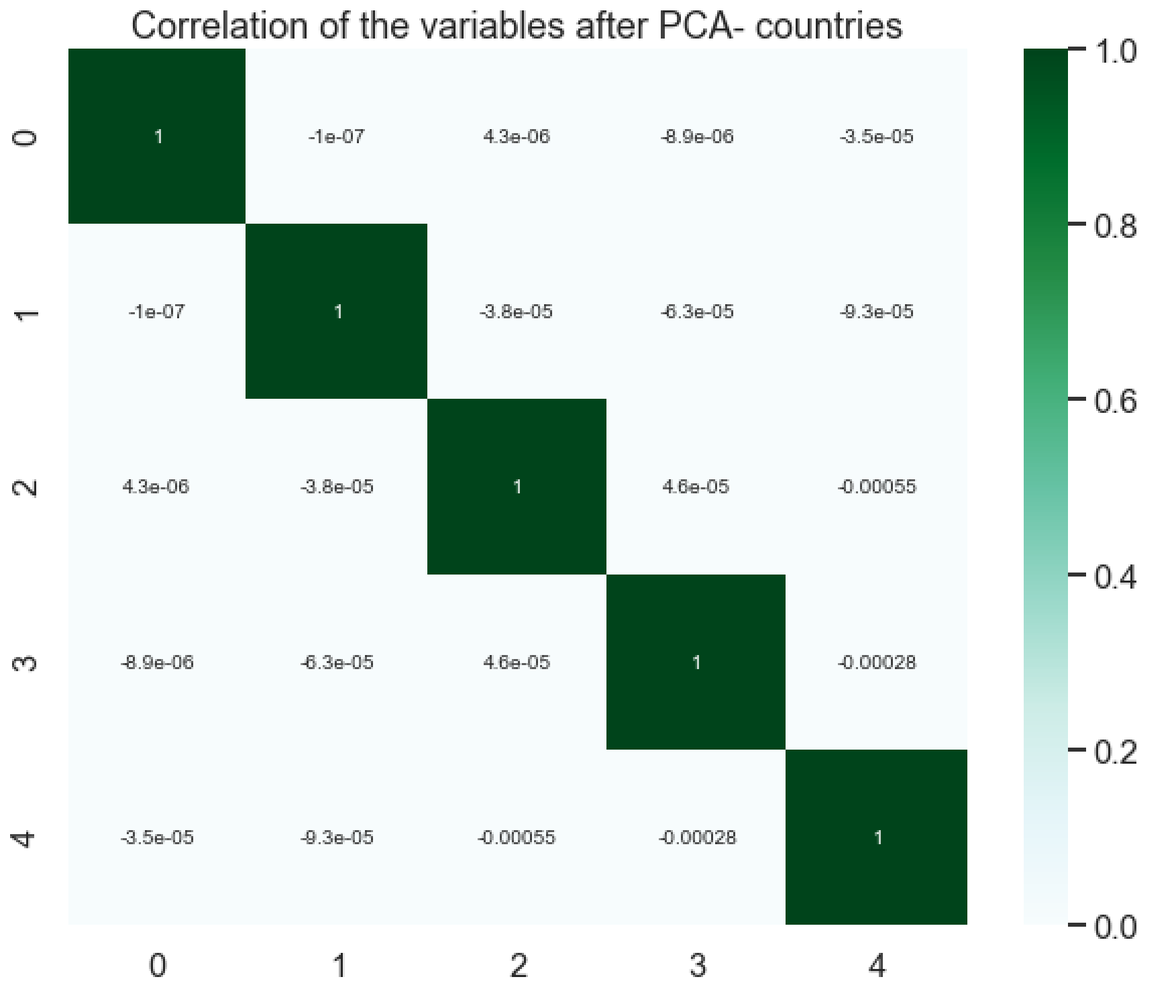




Correlation after PCA

INFERENCES

This shows that after performing PCA the multicollinearity has been removed as correlation is close to 0.



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 a 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 0.4111424002436614 4

For no. of cluster=4, silhouette score is 0.4120307839684570 5

For no. of cluster=5, silhouette score is 0.3018227058931638 5

For no. of cluster=6, silhouette score is 0.3760065208655269 4

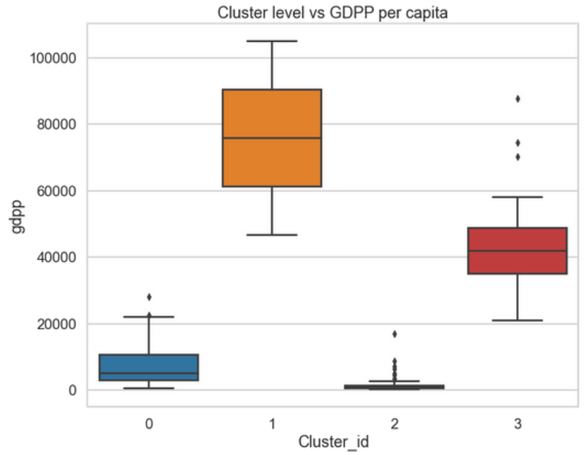
For no. of cluster=7, silhouette score is 0.2621318065720028

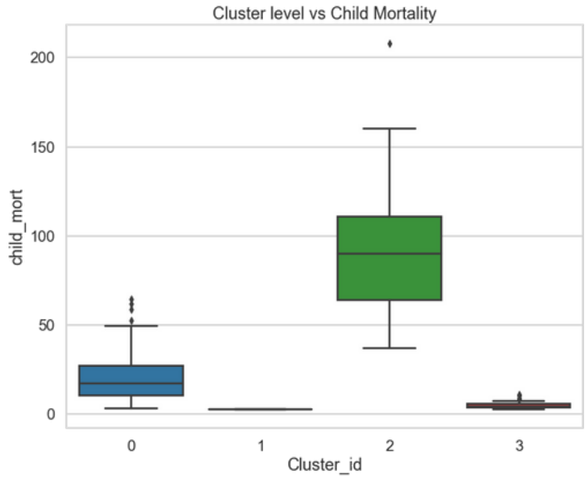
For no. of cluster=8, silhouette score is 0.2775830981601893

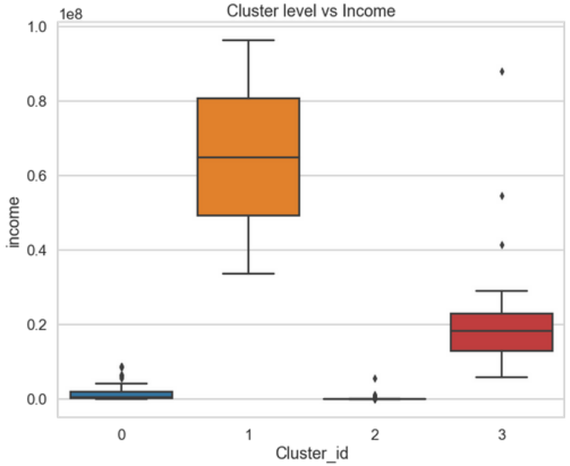
For no. of cluster=9, silhouette score is 0.2613300499433421 4

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

Visualization of the original variables with clusters







INFERENCES

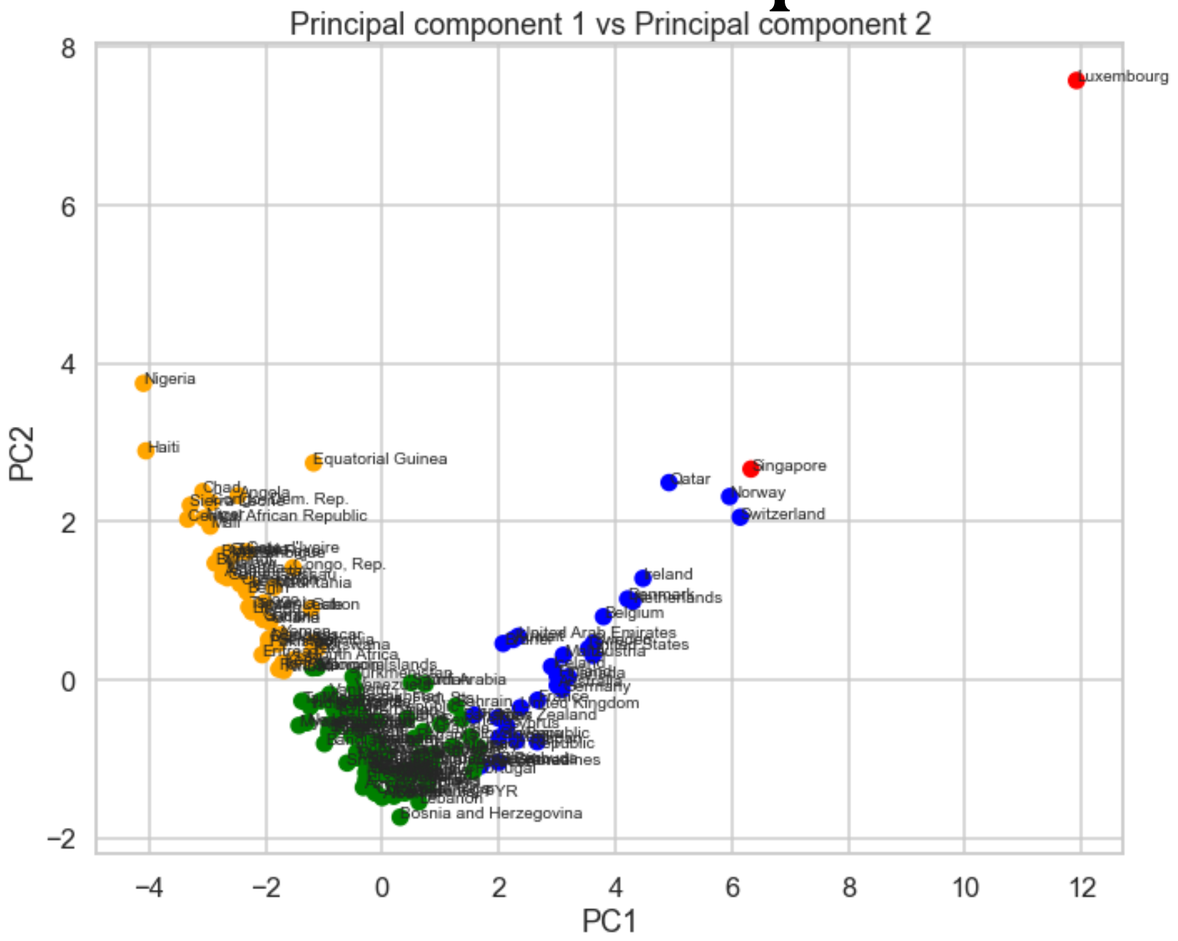
For cluster 0: values of gdpp and child mortality are slightly higher than the income with some outliers.

For cluster 1: values of gdpp and income are very high than other clusters, Mortality rate of children is very low than other clusters.

For cluster 2: values of gdpp and child mortality are very low while that of income are normally distributed and high.

For cluster 3: values of gdpp and income are normally distributed while that of child mortality is very

Visualization of principal component 1 and 2



INFERENCES

countries like 'Singapore'
and 'Luxembourg' are having high PC1 which means they are doing well,
while on other hand countries like 'Nigeria',
'Hiti', Equatorial Guniea and close to pc2 requires urgent need of aid.

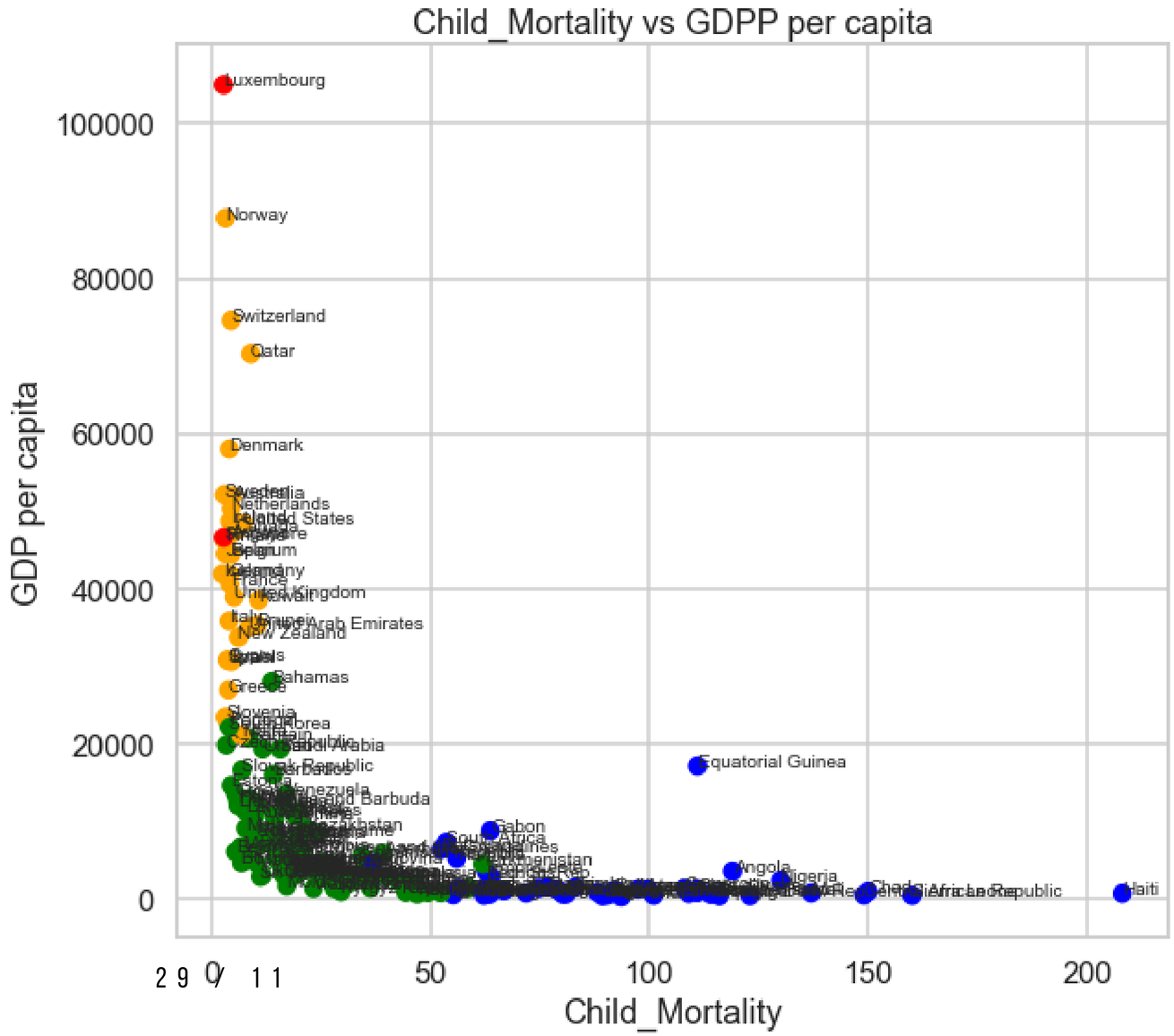
Visualizing (Child_mort vs gdpp) variables

INFERENCES

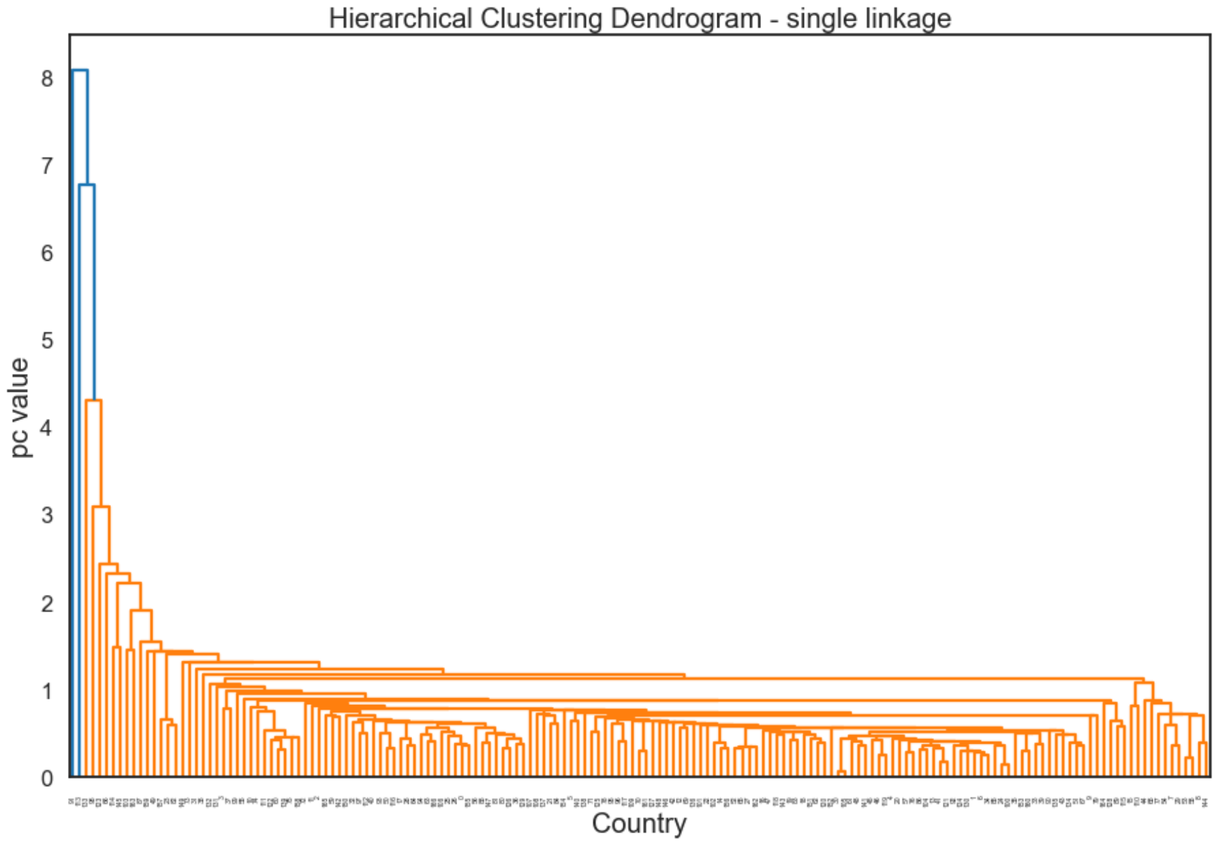
Country ‘Haiti’ is in dire need of aid.

Country ‘Luxembourg’ is having good gdpp and less child mortality rate.

‘Haiti’ and ‘Luxembourg’ are two outliers here as shown in the plot.



Hierarchical clustering (Linkage)



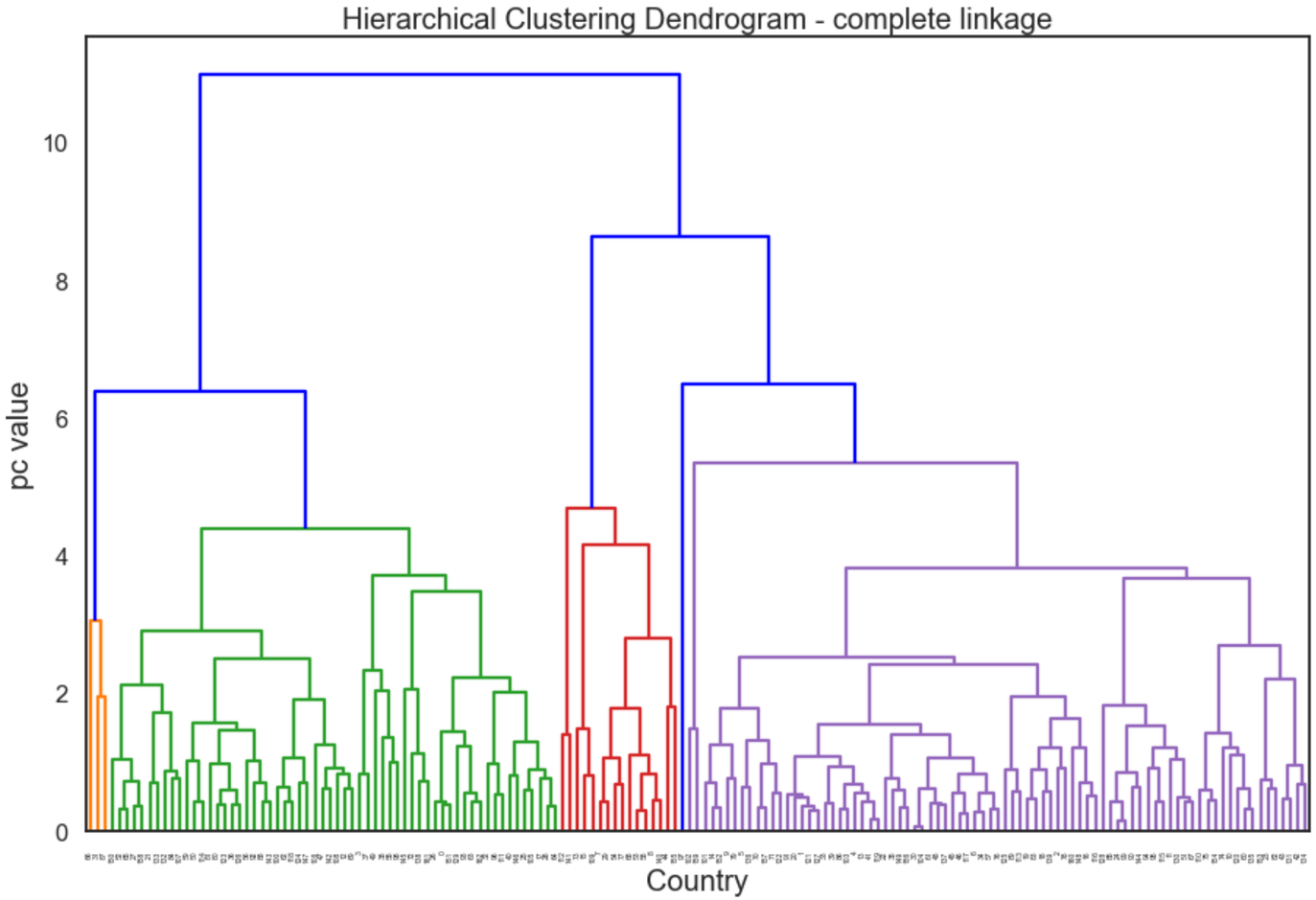
INFERENCES

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

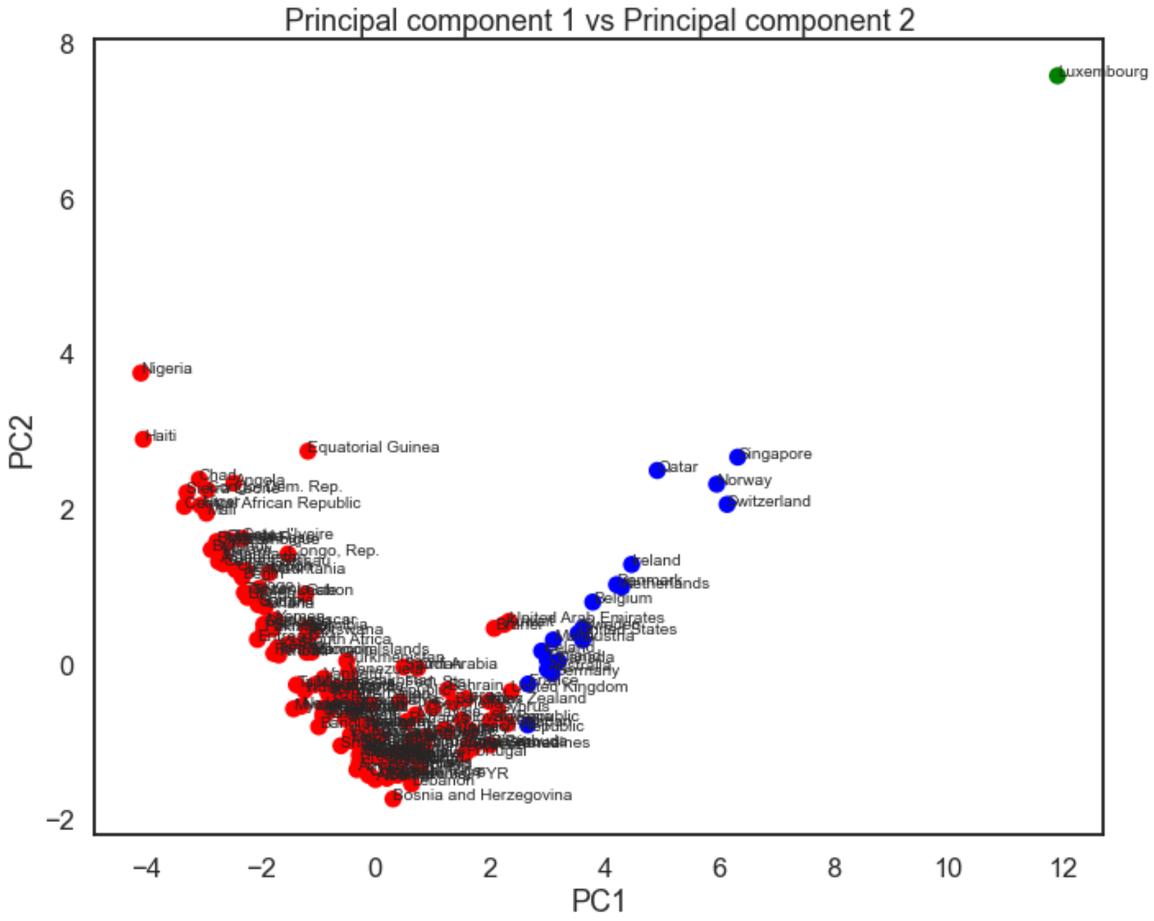
Hierarchical clustering (Complete Linkage)

INFERENCES

1. graph shows the proper way to decide the number of clusters at the threshold level. We will cut at 3 branches which will give us 3 clusters.



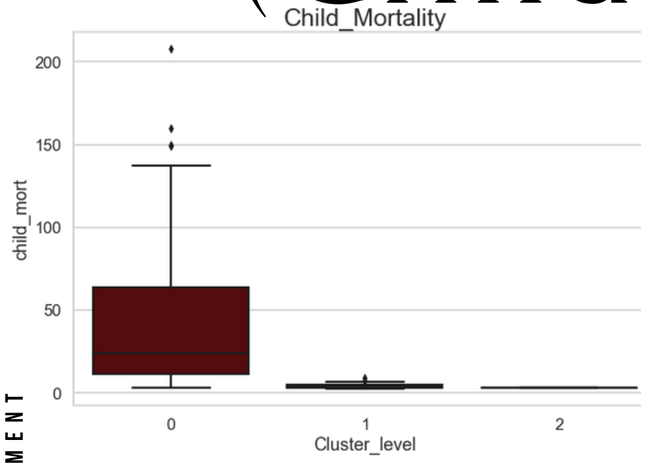
Visualization of hierarchical clustering

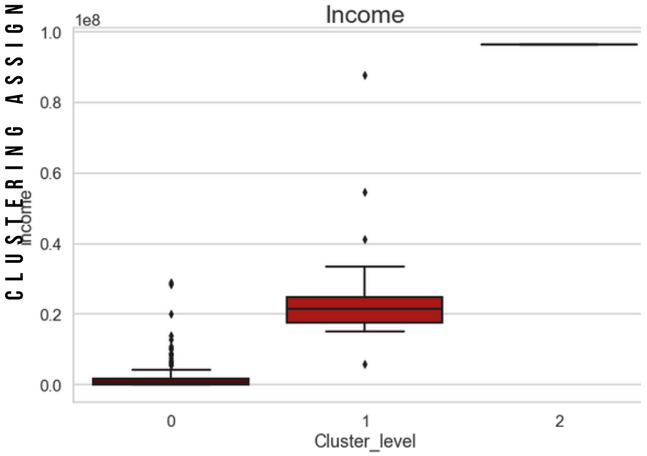


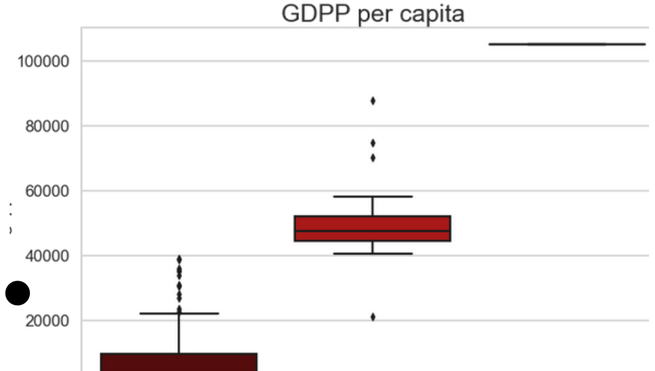
INFERENCES

1. The PC1 is in the direction where the countries need the least help.
2. Here, we are choosing PC1 because it has the maximum percentage of variance explained.
3. The countries with ‘Red’ colour data points need help in aid but the 'Blue' one does not require.

(Child mortality, Income and Gdpp)







INFERENCES

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

For cluster 1: values o income and gdpp are normally distributed.

For cluster 2: values of gdpp and income are high but values of Mortality of children is very less.

Insights (K-Means 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 in 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

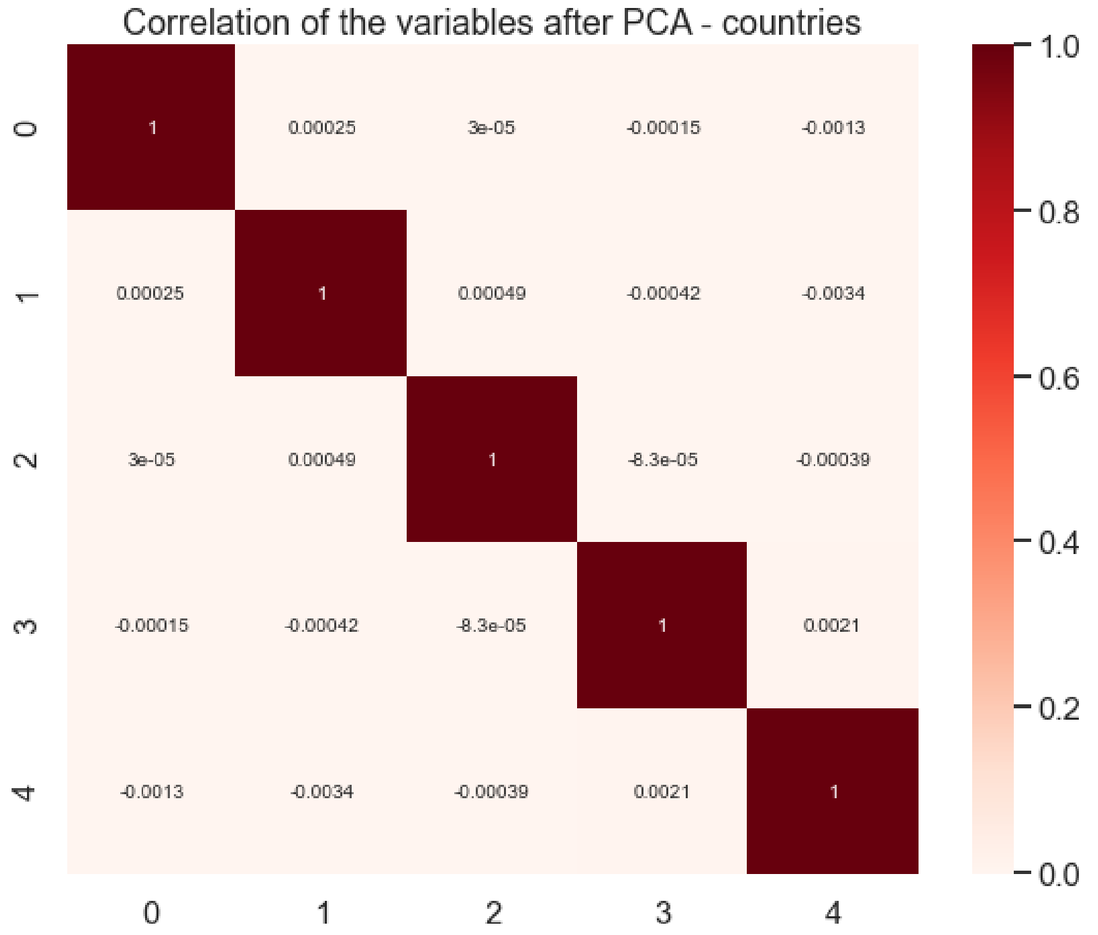
Hierarchical clustering :
Countries that are the direst need of aid Total 147 countries are in this category
Countries that are having good socio-economic and health factors
1 country is in this category – Luxembourg

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 more accurate output than hierarchical clustering. K-means gave precise information than hierarchical clustering.

Exclude outliers - K-Means clustering using PCA

INFERENCES

1. This shows that after performing PCA multicollinearity has been removed.

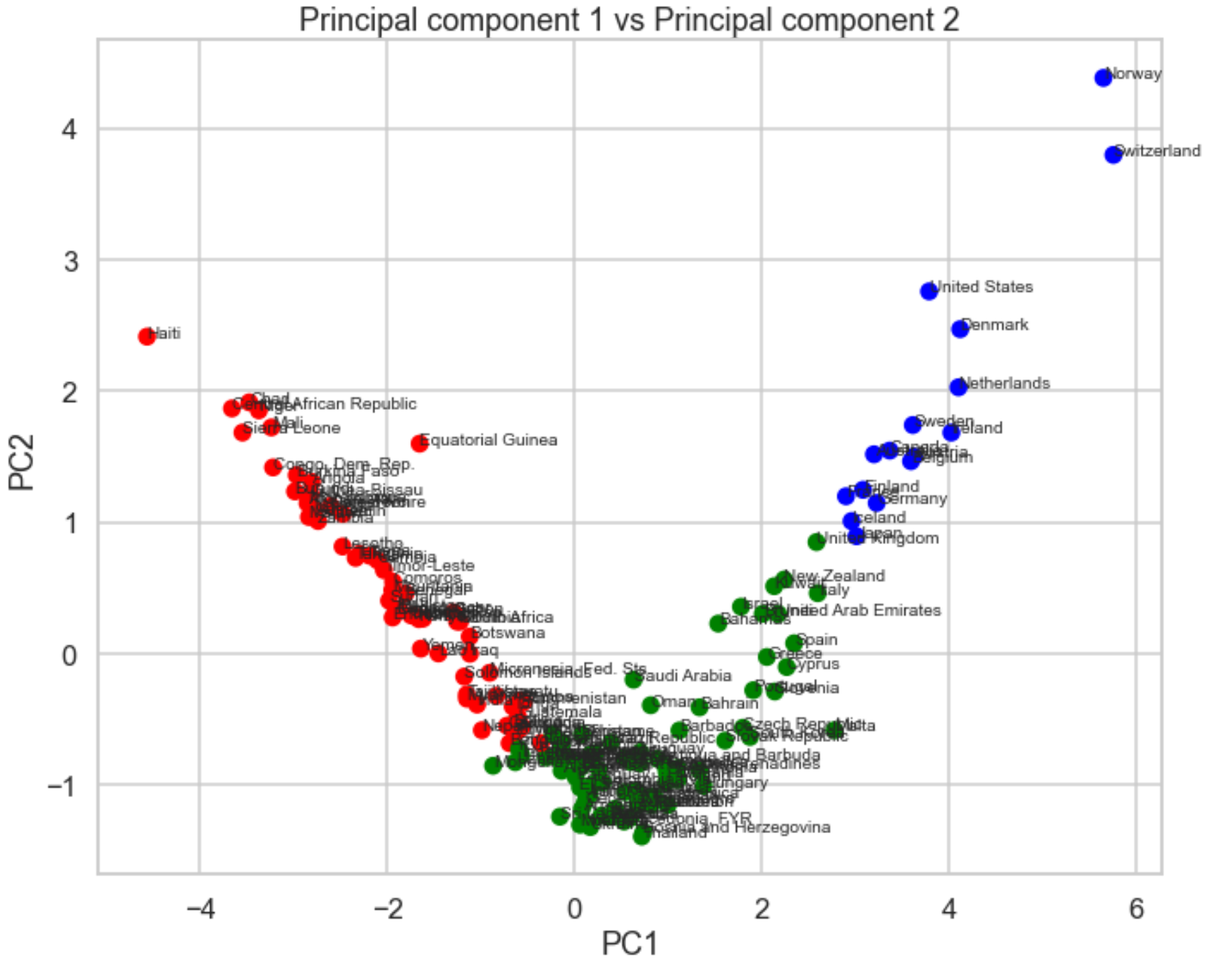


Visualization with PC1 and PC2

INFERENCES

1.The PC1 is in the direction where the countries are in need of least help. Here, we are choosing PC1 because it has maximum percentage of variance explained.

2. The countries with Red color datapoints are in urgent need of help\ aid



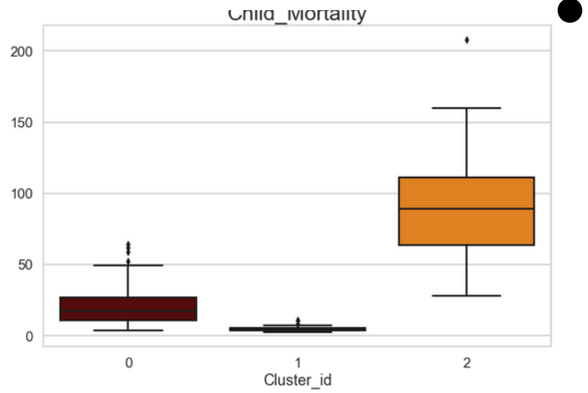
Visualization of variables (gdpp, income and child mortality)

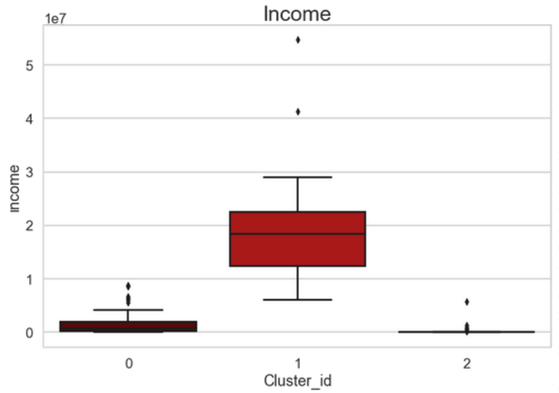
INFERENCES

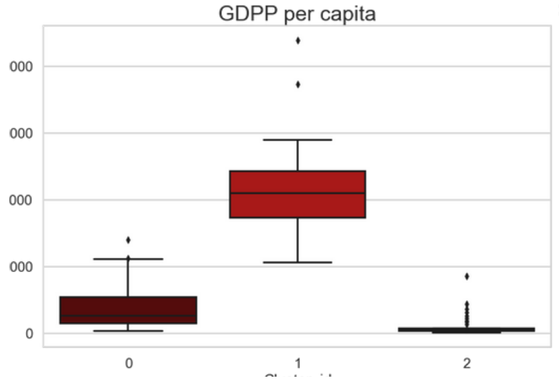
1 . For cluster 0: values of
all the variables are normally distributed.

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.







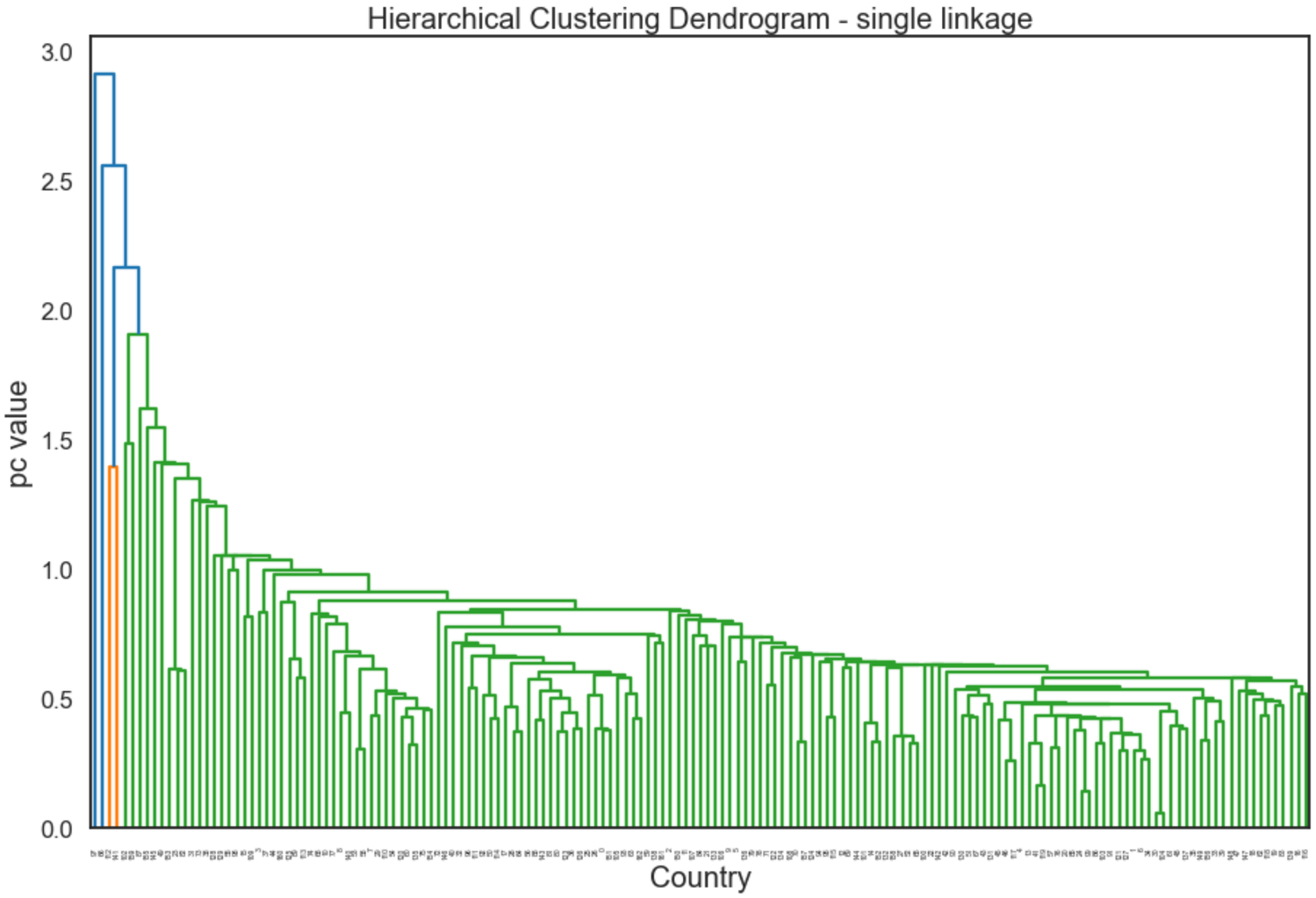
Insights (K-Means exclude outlier)

INFERENCES

1.There are a total of 47 countries from the dataset in need of urgent help/aid as they are having the lowest income, high child mortality and low GDP per capita.

28 countries are having good socio-economic and health factors.

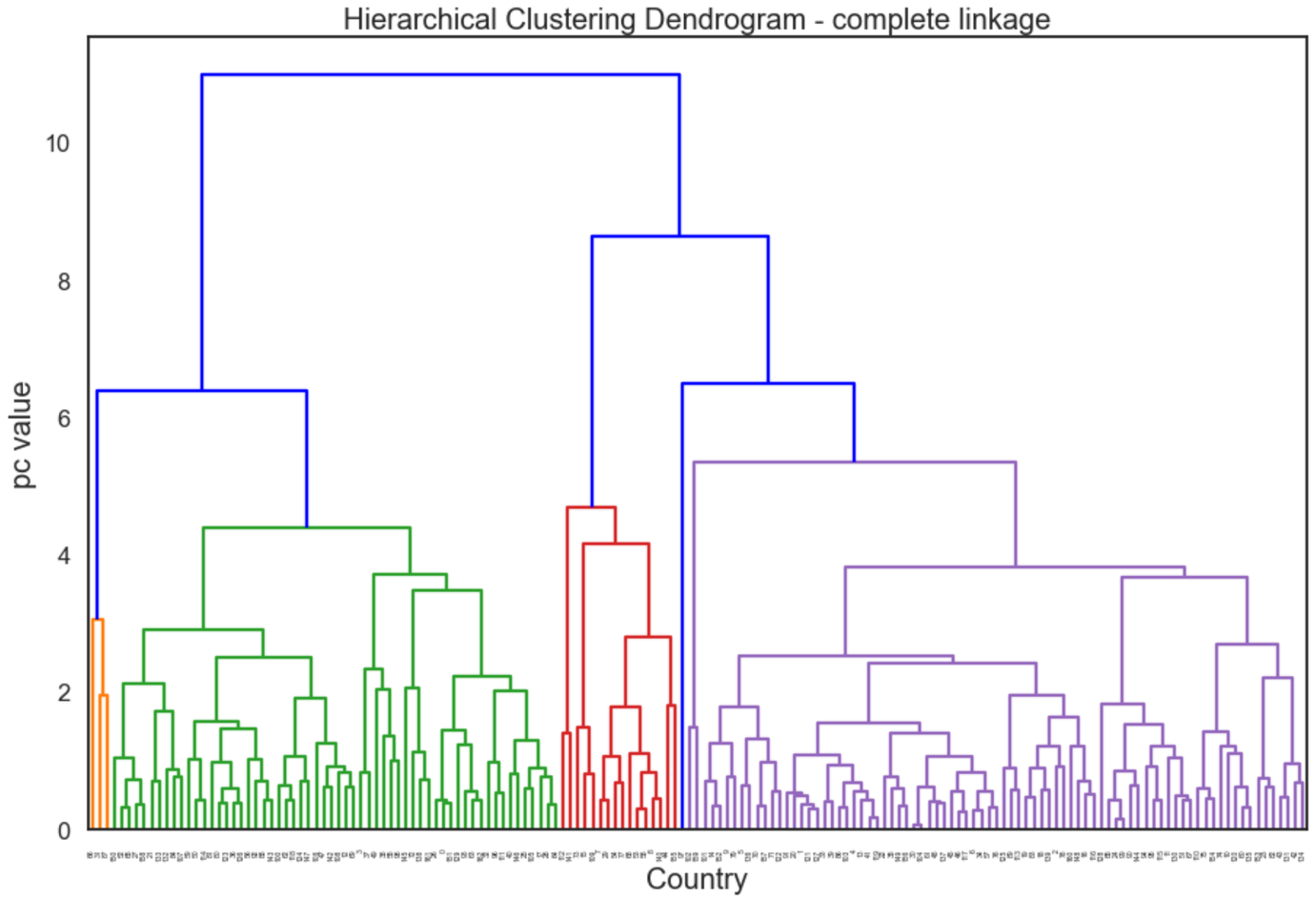
Hierarchical clustering (Linkage) – excluding outliers



INFERENCES

1. As we can see from the graph of linkage dendrogram, it is not quite visible and doesn't 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) – excluding outliers

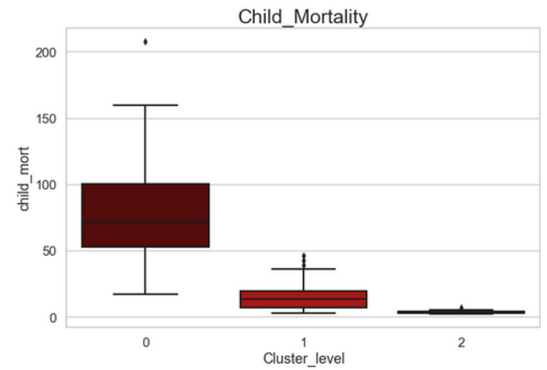


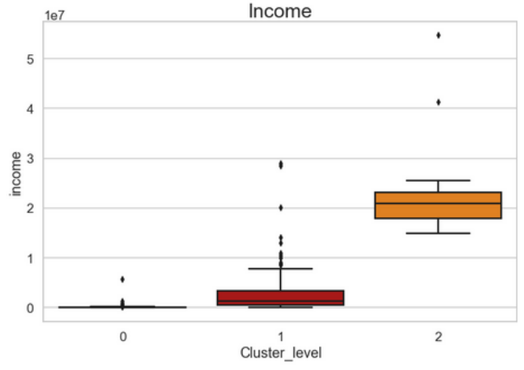
INFERENCES

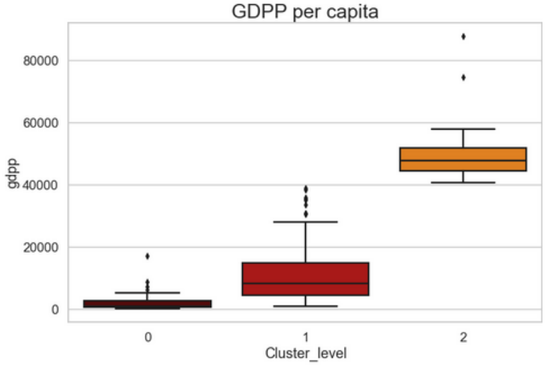
1. This graph shows the proper way to decide number of clusters needs to be used by cutting at threshold value.

2. We will cut at 3 branches which will give us 3 clusters

(Child mortality, Income and Gdpp)







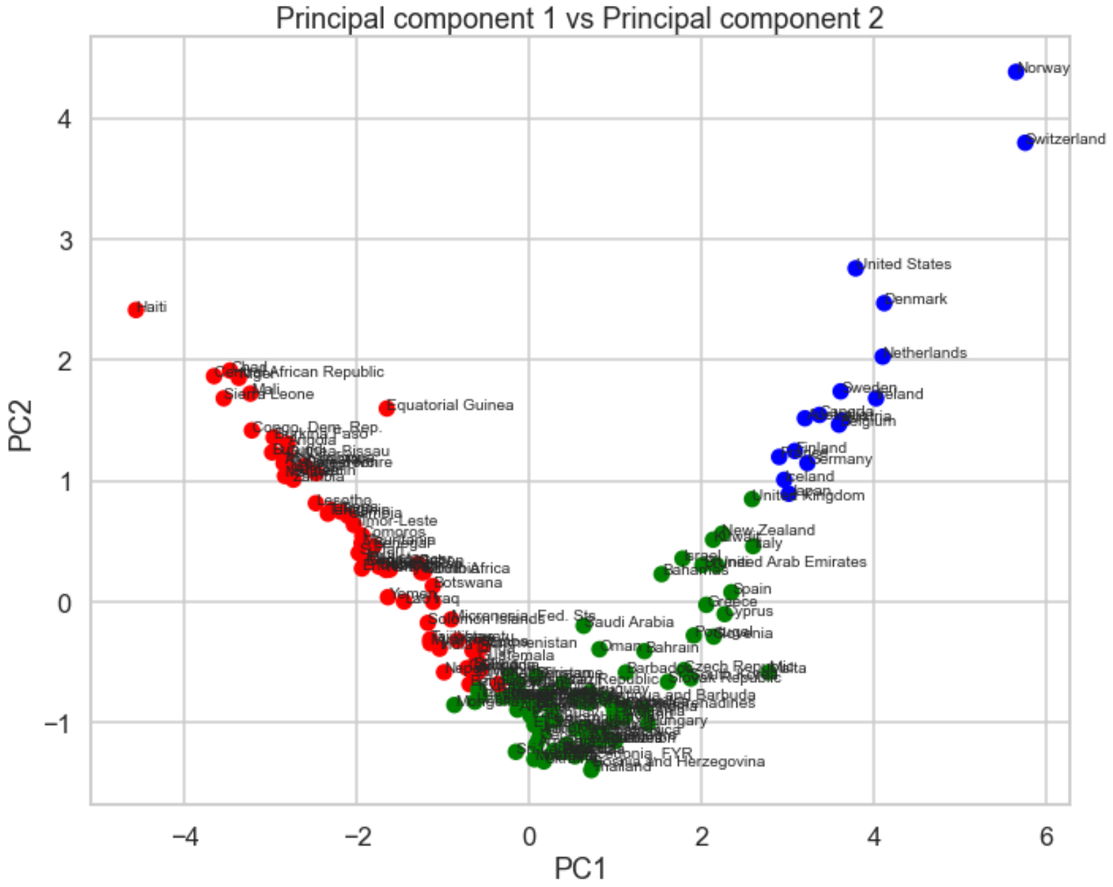
INFERENCES

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

2. 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

3. 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 – excluding outlier



INFERENCES

1. the PC1 is in the direction where the countries need of least help. Here, why we are choosing PC1 because it has the maximum percentage of variance explained.

2. The ‘Red’ colour 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 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

Conclusion – Without outlier

K-Means vs Hierarchical Clustering

K-means clustering :

Countries that are in direst need of aid Total 47 countries are in this category.

Hierarchical clustering : Countries that are in 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 KMeans 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!