1. Problem Statement:
With a corpus of $10 million, our goal is to provide financial assistance to the countries which are in the direst need of aid. This money will be used for the overall development of these countries and to deal with issues such as providing good healthcare so as to reduce the number of deaths due to medical negligence, to provide the children and expecting mothers with proper nutrition in their infant years / pregnancy duration so as to reduce the child mortality rate, generating employment in the small scale / cottage industries to increase the per capita income and at the same time the GDP per capita of the country, etc.

2. Data:
The data being used to filter out the countries in need of aid consists of the following metrics:

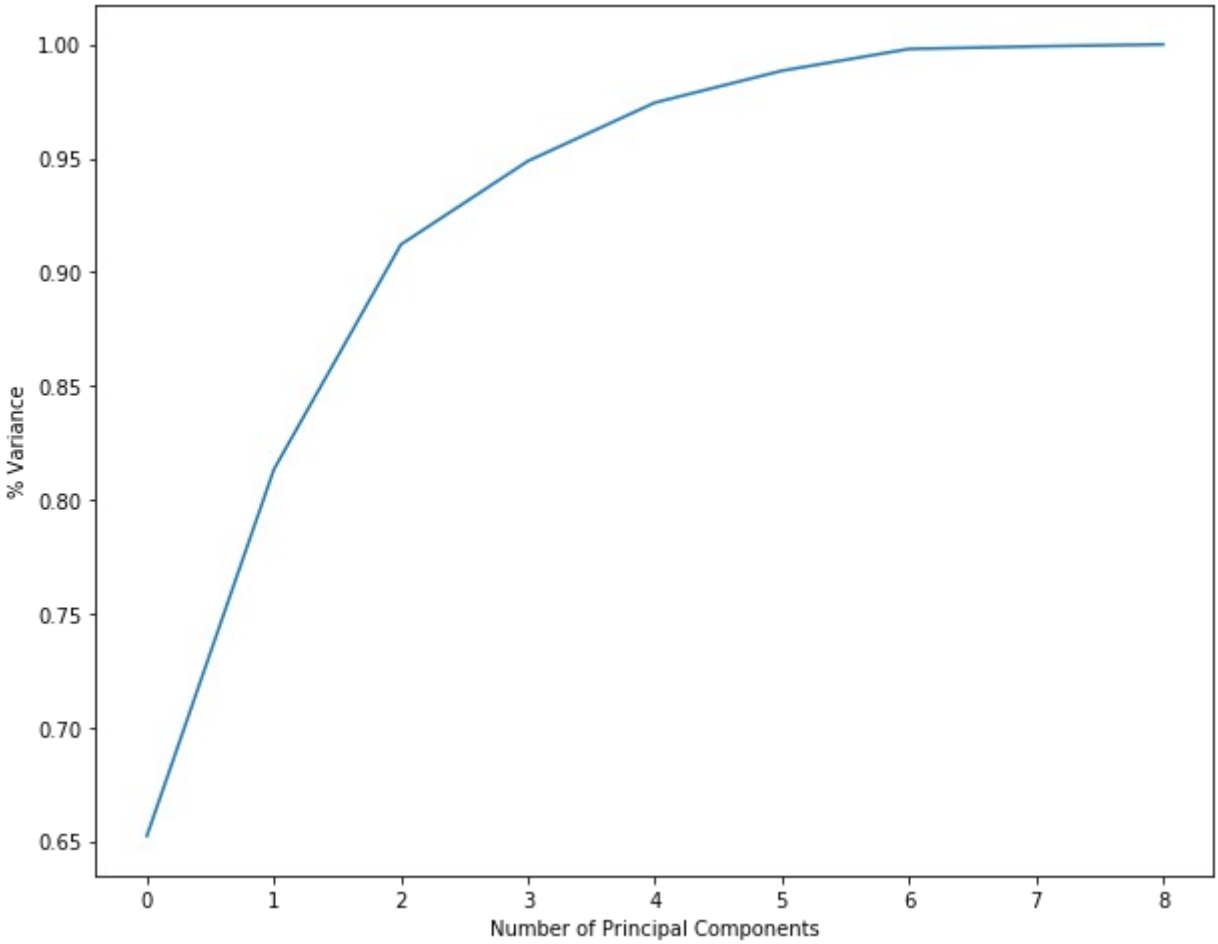
Metric	**Description**
Child Mortality	Number of deaths per 1000 live births for children under 5 years of age.
Income	Net income per person
GDPP	GDP Per Capita
Exports	Exports of goods and services done by the country as a percentage of GDP.
Imports	Goods and Services imported by the country as a percentage of GDP
Health	Money spent on health as a a percentage of GDP
Inflation	Measurement of the total growth rate of the GDP
Life Expectancy	Average life expectancy for the country
Total Fertility	Average number of children born to a single woman.

3. Analysis Approach:
The heart of the analysis lies in making the data ready for clustering, so that the countries with similar characteristics can be grouped together in a single bucket.
The columns given as a percentage of the total GDP are first converted into their actual values since by intuition we know that there is a huge gap between the rich and the poor countries and a representation of key factors in percentage of the total GDP might be misleading.

4. Preparing the data:
The data is then treated for outliers keeping in mind the end goal of the analysis i.e filtering out countries in the direst need of financial aid. With this in mind, countries with extremely high values of income per person and extremely low child mortality rate have been filtered out. This leaves us with a stable dataset on which technical analysis will not be misleading.

5. Principal Component Analysis:
Since a model has to built on the dataset, a less number of variables in the model will render it more stable, and to this effect Principal Component Analysis has been used to capture the maximum information possible and at the same time reduce the dimensionality of the data. The data has been standardized and then the required number of Principal components have been used to build the model. This is done by making a Scree plot.

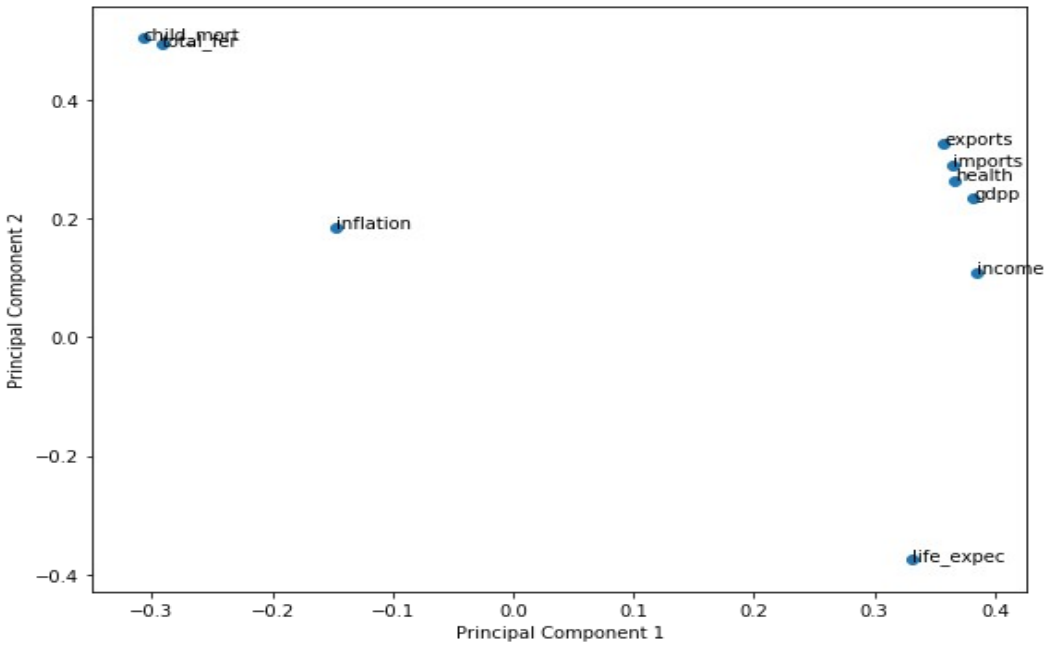
Below is the Scree plot obtained.



It is evident from the above plot that 3 Principal componets are able to capture around 90% of the variance whereas 4 Principal components will capture about 95% of the variance.

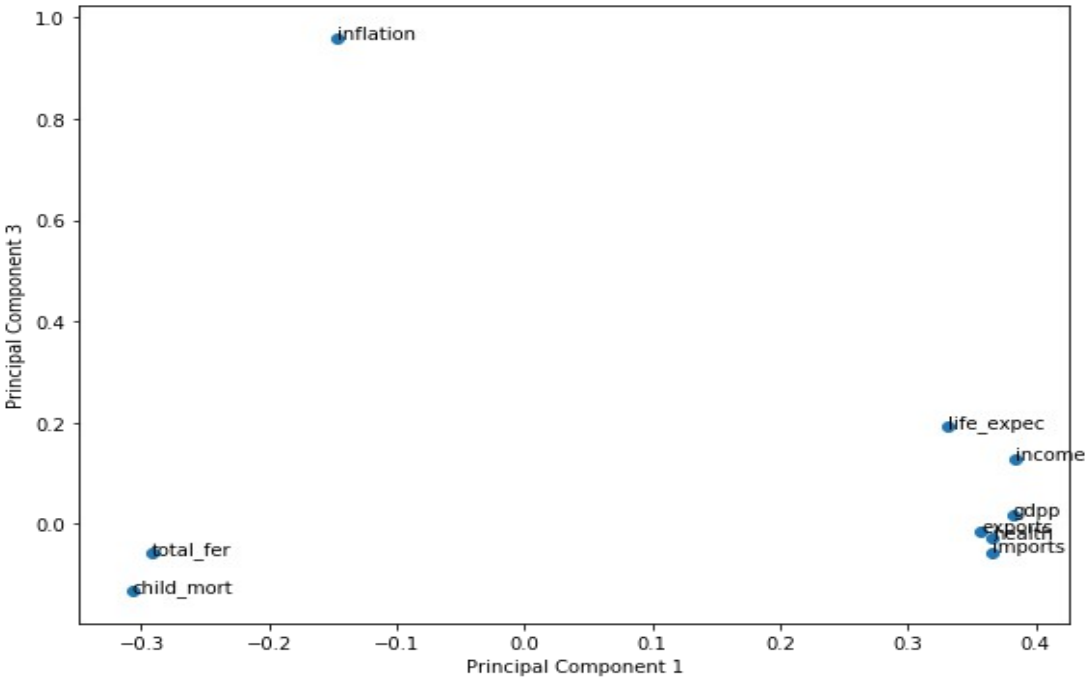
Three Principal Components have been chosen for this analysis.

The scatter plot between these componets shows us the weightage given to each of the variable. *Below is the plot between PC1 and PC2.*



It can be seen from the above plot that PC1 tends towards a high rate of child mortality and fertitlity whereas PC2 tends towards a greater life expectancy and a higher net income per person.

Below is the scatter plot between PC1 and PC3.



It can be seen from the above plot that PC3 tends towards a lower fertility rate and child mortality rate whereas the inflation is high.

6. Model Building:

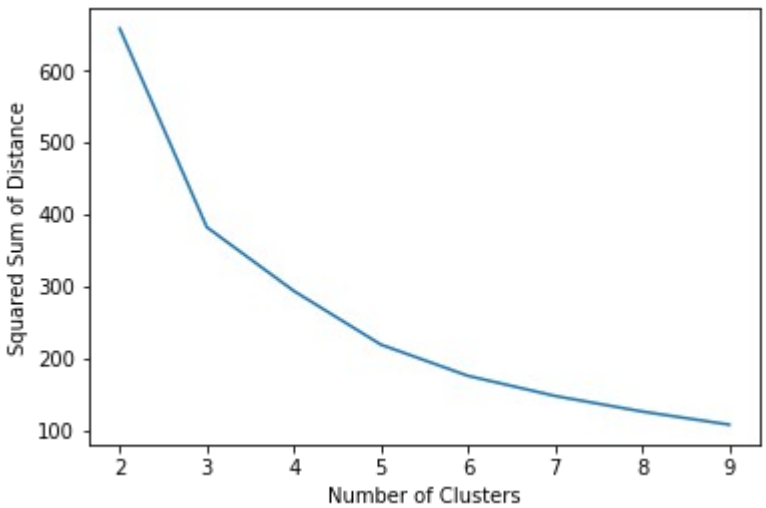
Both K-Means and Heirarchical Clustering methods have been used to try and group the countries.

6.1 K-Means Clustering:

The required number of clusters are found by using the Elbow-Curve and the Silhouette Score methods. The final number of clusters has been decided by considering both the methods as well as the practical nature of the task at hand.

6.1.1 Elbow-Curve Method:

The below curve is obtained.



It is seen that there is a sharp decrease in the Squared Sum of Distance (SSD) when the number of clusters is increased from 2 to 3. The second elbow is seen when the number of clusters is increased to 5. After that the decrease in SSD is really low as the number of clusters are increased.

6.1.2 Silhouette Score Method:

The following score is obtained for the given number of clusters:

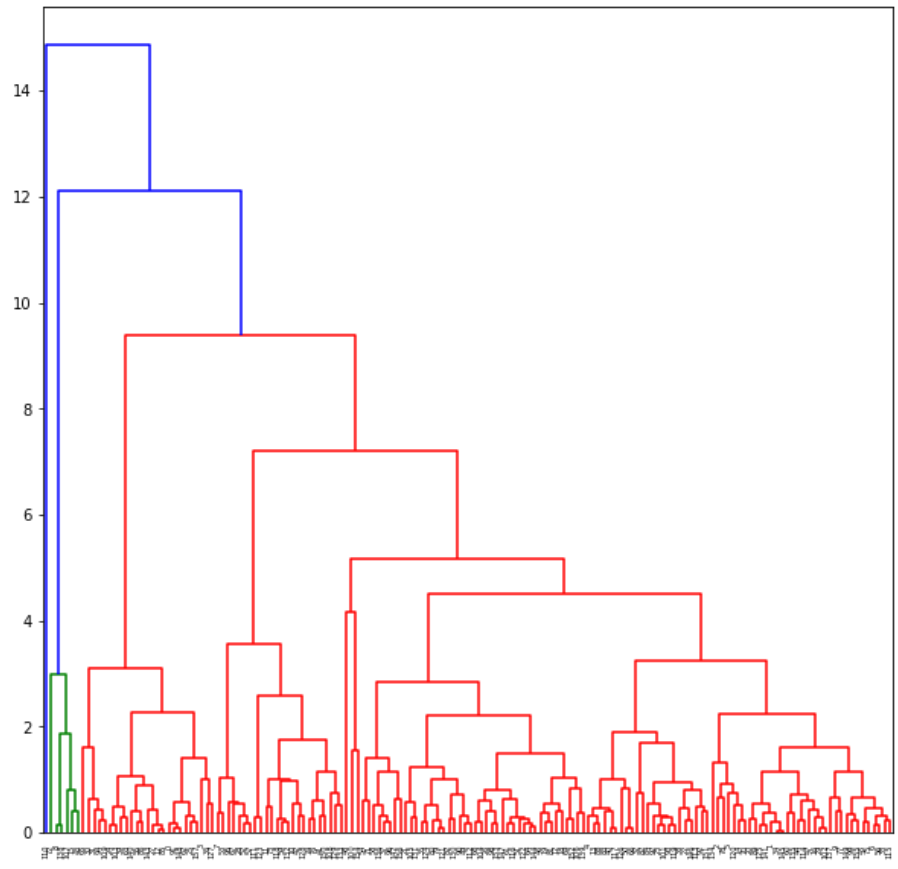
Number of Clusters	Silhouette Score
2	0.53
3	0.49
4	0.46
5	0.47
6	0.40
7	0.41
8	0.37
9	0.36

It can be seen that the Silhouette Score is the highest for 2 clusters and then starts decreasing as the number of clusters increases.

From the elbow-curve method, silhouette score analysis and the practical aspects while grouping countries, the total number of clusters has been taken as 3.

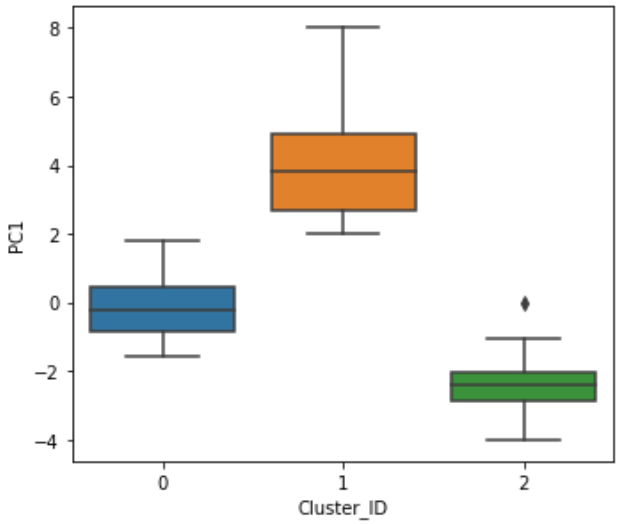
6.2 Heirarchical Clustering:

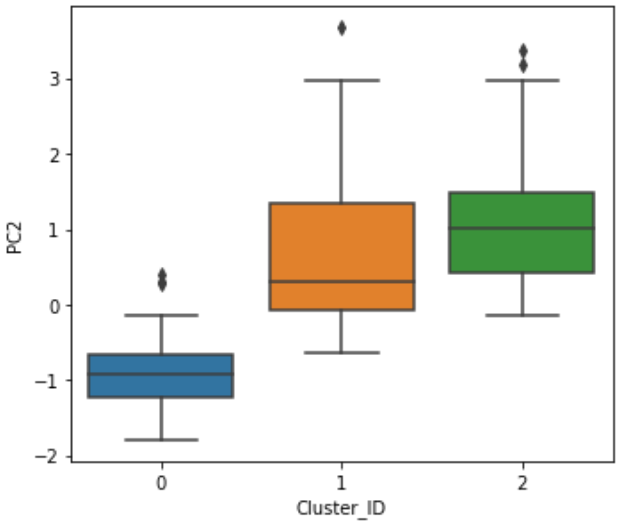
All three types of linkages have been tried for clustering, however only complete linkage had a distinguishable dendrogram. *Below is the dendrogram obatined.*

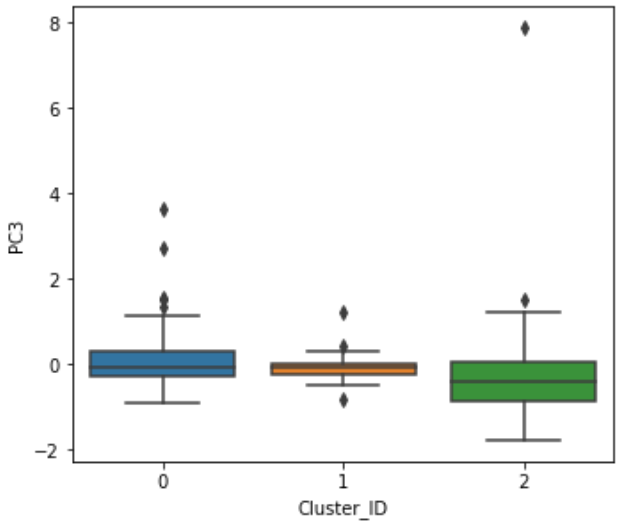


It can be seen that for 3 clusters, the heirarchical clustering is not feasible since most of the countries will be grouped under 1 cluster. We can try increasing the number of clusters for a more balanced distribution of countries, however, the practical aspects prohibit us from doing so since the final clusters obtained would be too complicated to understand.

Therefore, we will proceed with the clusters obtained via the K-Means Clustering. The distribution of each of the Principal Components for each cluster is as follows:





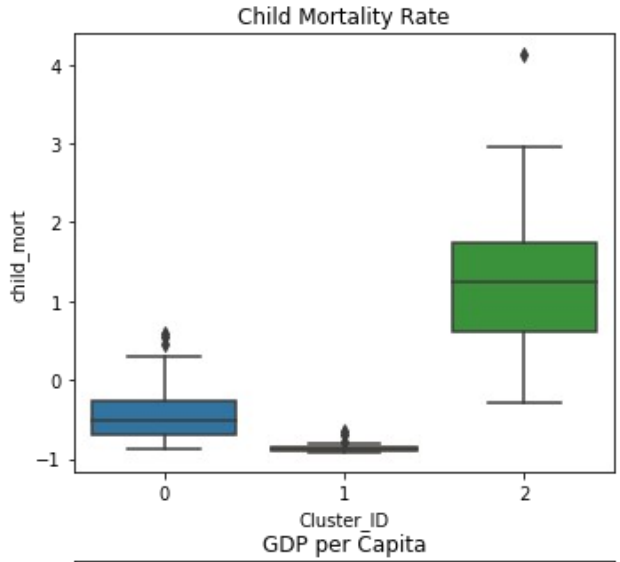


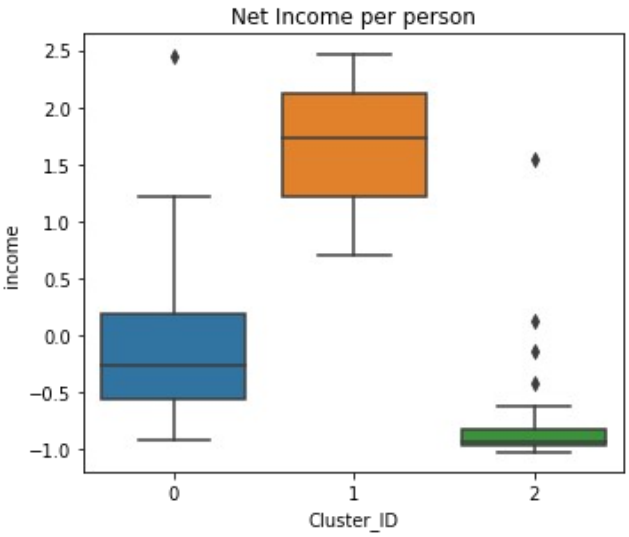
It can be seen that the clusters are neatly distinguishable.

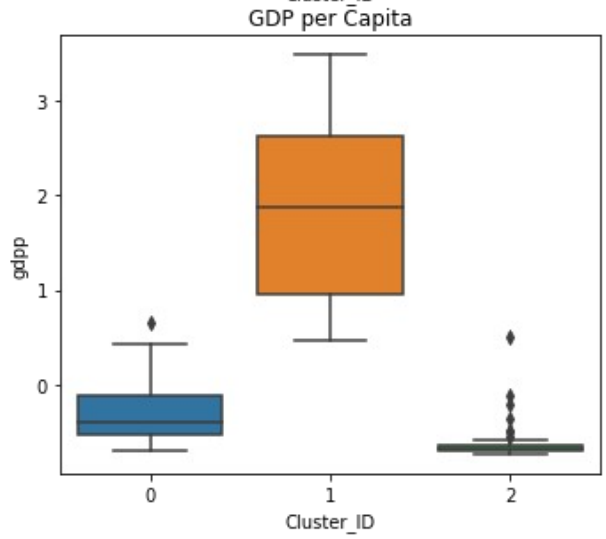
As mentioned above, three clusters have been used for grouping the countries. The grouping gives us the following distribution.

Number of Countries	**Cluster ID**
83	Developing Countries
29	Developed Countries
46	Under developed Countries
158	**Total**

Below is the distribution of the Child mortality rate, Net income per person and the GDP per capita for each of the clusters.







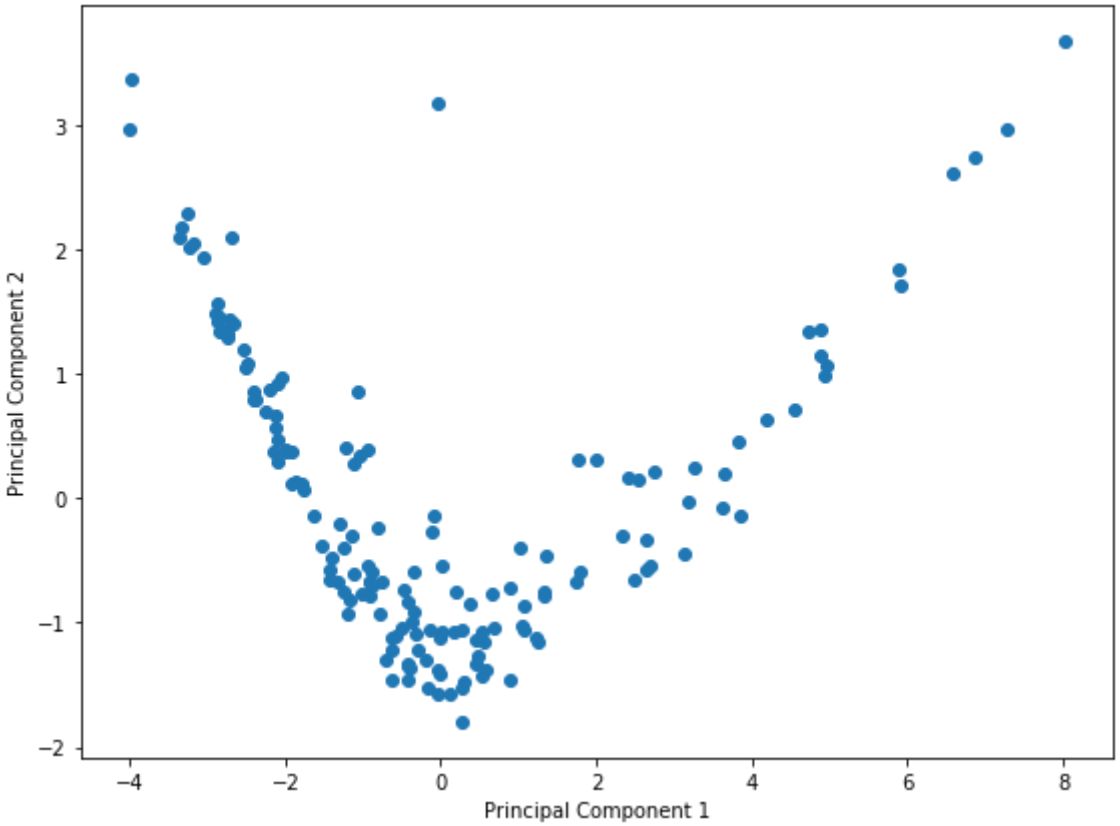
Interpreting the results:
Cluster 2 represents the under-developed countries since they have a really high child mortality rate combined with a lower net income per person and a lower GDP per capita. These are the countries that need assistance to overcome their problems since they are not self sufficient in terms of financial resources.

Cluster 1 represents the developed countries which enjoy a high income, higher GDP per capita and a really low child mortality rate.

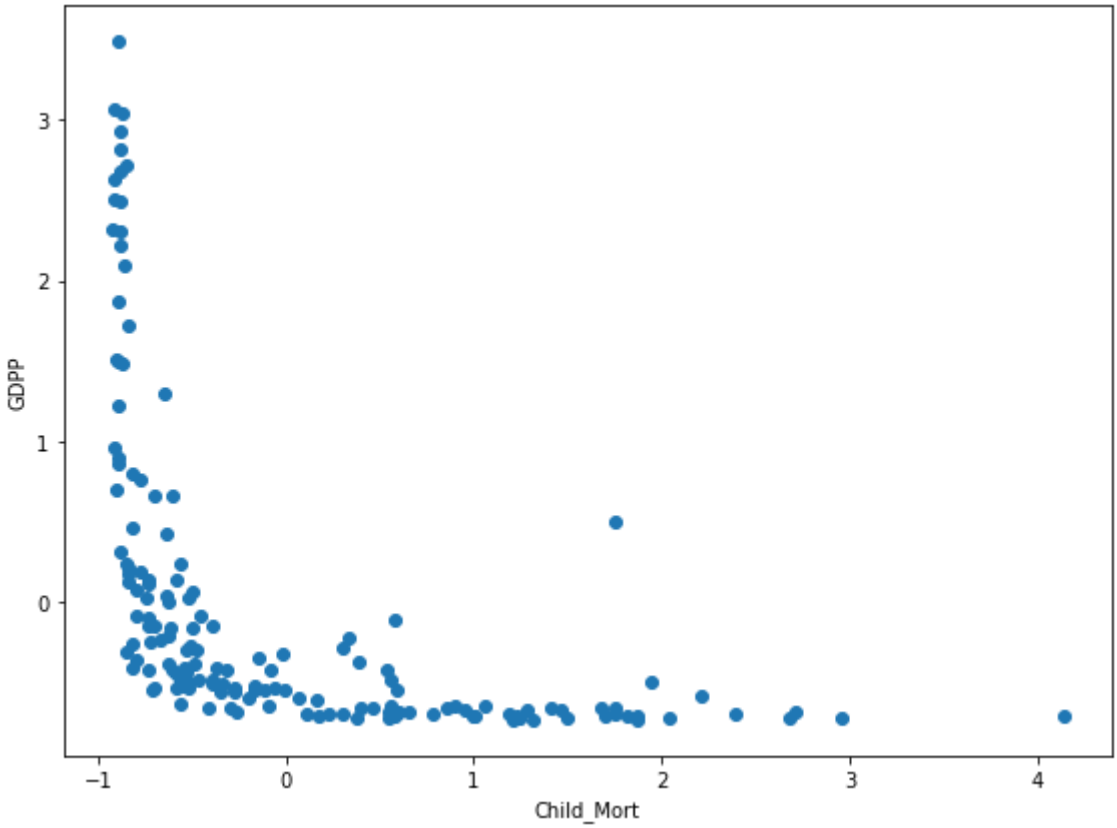
Cluster 0 represents the developing countries since they have a medium rate of child mortality, also their net income per person is significantly higher than the poor countries but at the same time significantly lower than the developed countries. The GDP per capita for these coutnries also follow the same pattern, i.e higher than the underdeveloped countries and lower than the developed countries.

To get a more wholesome view of the clusters formed, below are some plots that clearly outline the various groups of countries.

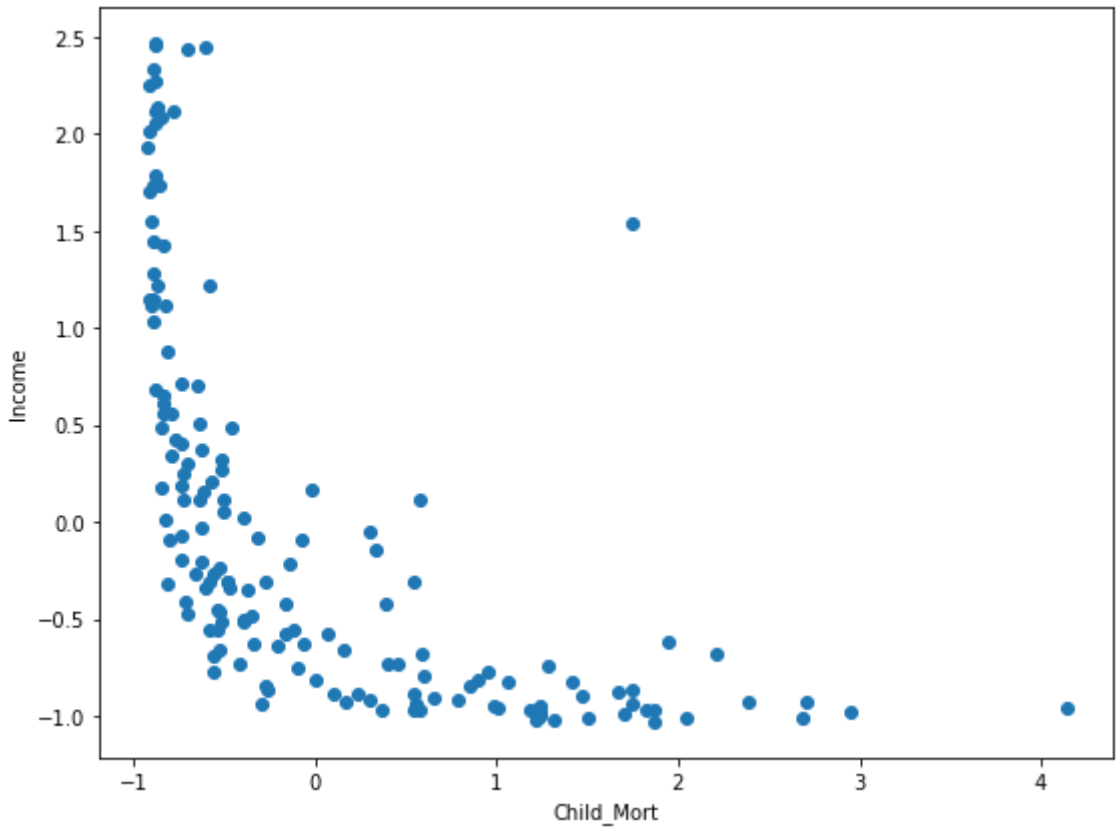
Scatter plot between Principal Component 1 and 2.



Scatter plot between GDP Per Capita and the Child Mortality Rate:



Scatter plot between Net Income Per Person Vs Child Mortality Rate:



Conclusion:
The underdeveloped countries are futher sorted by their child mortality rate, net income per person and the GDP per capita and the below 5 countries have been found to be in the most need of assistance.

S.No.	**Country**
1.	Haiti
2.	Seirra Leone
3.	Chad
4.	Central African Republic
5.	Mali

Below is also a list of all the 46 underdeveloped countries in alpabetical order that need financial assistance at some level in order to improve the living conditions of their population.

S.No.	**Country**
1	Afghanistan
2	Angola
3	Benin
4	Burkina Faso

S.No.	Country
5	Burundi
6	Cameroon
7	Central African Republic
8	Chad
9	Comoros
10	Congo, Dem. Rep.
11	Congo, Rep.
12	Cote d'Ivoire
13	Equatorial Guinea
14	Eritrea
15	Gabon
16	Gambia
17	Ghana
18	Guinea
19	Guinea-Bissau
20	Haiti
21	Kenya
22	Kiribati
23	Lao
24	Lesotho
25	Liberia
26	Madagascar
27	Malawi
28	Mali
29	Mauritania
30	Mozambique
31	Namibia
32	Niger
33	Nigeria
34	Pakistan
35	Rwanda
36	Senegal
37	Sierra Leone
38	Solomon Islands
39	South Africa
40	Sudan
41	Tanzania
42	Timor-Leste
43	Togo
44	Uganda
45	Yemen
46	Zambia