Clustering and PCA Assignment 1

Problem Statement

HELP International is an international humanitarian NGO that is committed to fighting poverty and providing the people of backward countries with basic amenities and relief during the time of disasters and natural calamities. It runs a lot of operational projects from time to time along with advocacy drives to raise awareness as well as for funding purposes.

After the recent funding programmes, they have been able to raise around $ 10 million. Now the CEO of the NGO needs to decide how to use this money strategically and effectively. The significant issues that come while making this decision are mostly related to choosing the countries that are in the direst need of aid.

And this is where you come in as a data analyst. Your job is to categorise the countries using some socio-economic and health factors that determine the overall development of the country. Then you need to suggest the countries which the CEO needs to focus on the most.

Solution:

The approach I am trying to follow to find the solution to the above problem is to try segregating the countries into different clusters using both k-means and hierarchical clustering algorithms and then analysing those clusters to identify the countries which are in direst need of the above aid.

I will also make use of Principal Component analysis(PCA) to reduce the dimensions of the overall dataset and come to a feasible number of non collinear principal components on which the original data set can be projected and then clustering algorithms can be applied.

Data reading, understanding and cleaning

I started with reading and understanding the data by importing the dataset into a dataframe. I looked for areas where data cleaning may be required, but data seemed to be appropriate and didn't require any data cleaning.

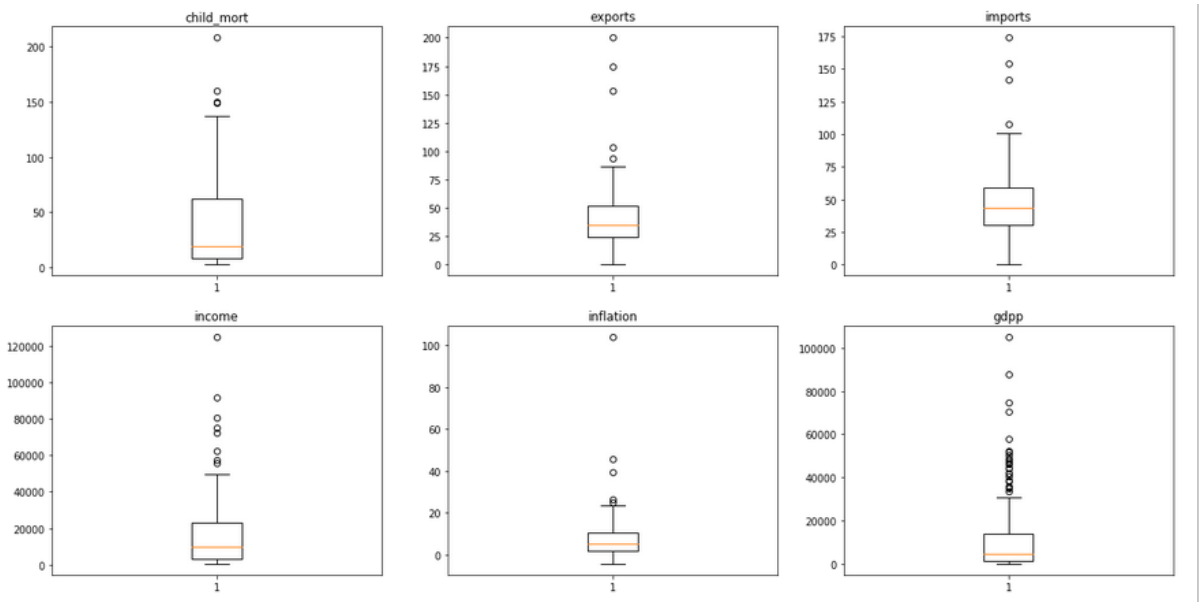
Data preparation - Outlier analysis & Standardization

I then moved onto the outlier analysis which is essential to be performed before the PCA as it may affect the outcomes of the PCA.

While doing the outlier analysis, I found there were indeed outliers (as seen from figure below) for variables such as child_mort, exports, imports, income, inflation and gdpp.

Hence I decided to remove some of these outliers from the dataset before doing the clustering. However not all outliers were removed so as not to delete too much information from the dataset.

After removing the outliers, I lost less than 10% of the original dataset and that seems okay.



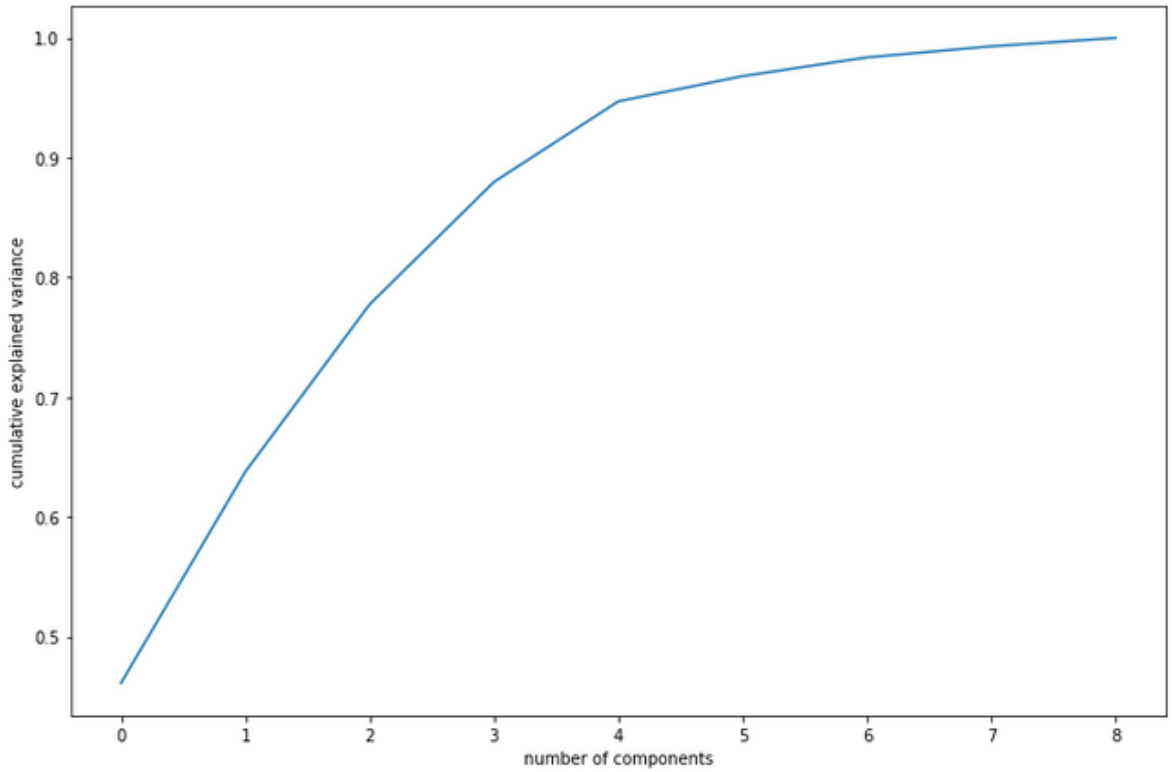
After removing the outliers, I standardized the data to get all the variables into a same scale which is another important pre-requisite before performing the PCA.

Principal Component Analysis

PCA was performed on the standardized dataset using the PCA module of the sklearn library.

To come up to a viable number of components, scree plot was plotted to check the cumulative variance that can be explained by the number of principal components.

The scree plot is shown below:

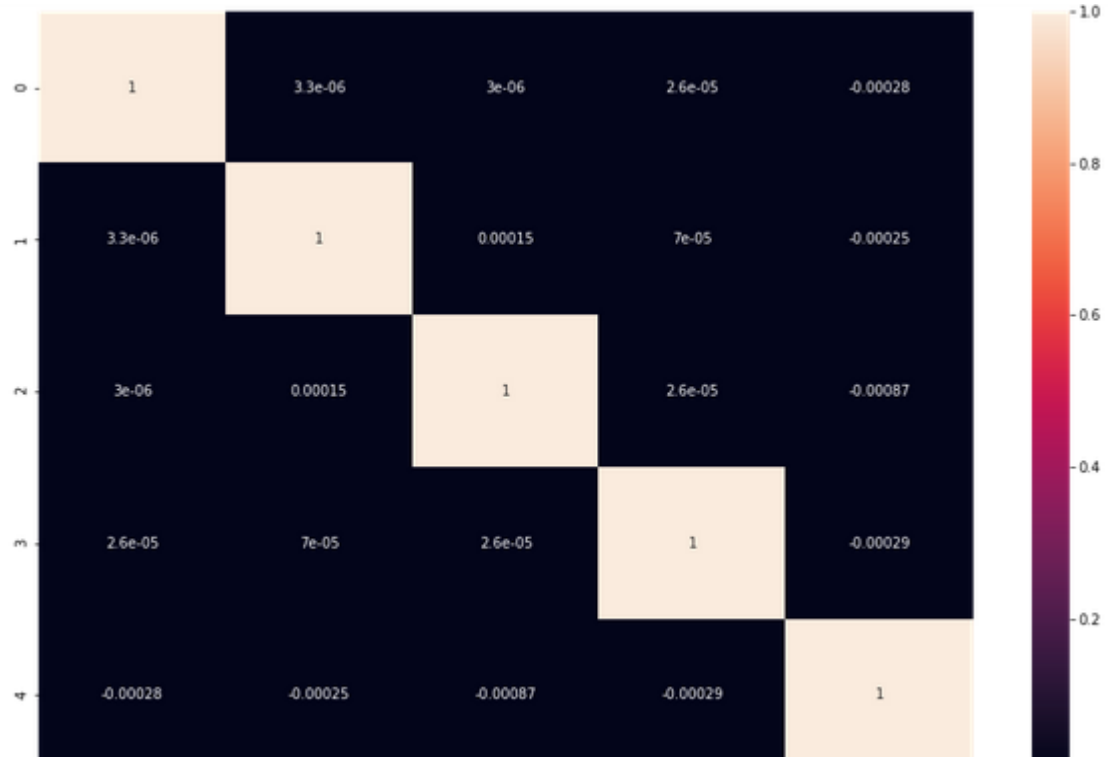


From above, we can see that PC = 5 explains more than 95% of the variance in the data and hence we'll choose 5 as the number of PC's for our modelling.

Incremental PCA was used to finalize these 5 principal components.

And then I did the basis transformation of the original data onto these 5 PC's

After doing the basis transformation, we checked the correlation between these 5 PC's and found there were no correlation between them at all as seen below.

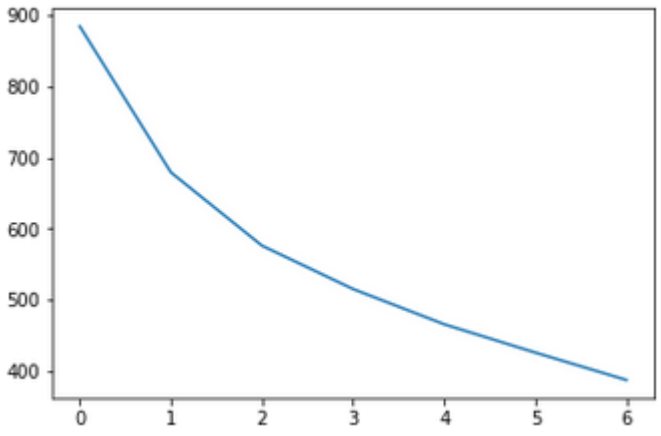


K-means Clustering

After performing the PCA, it was time to perform the k-means clustering on this modified dataset.

I started with an arbitrary k = 4 and found the cluster labels. However to come to an optimal number clusters, I checked the **elbow curve and did the silhouette analysis to come to a viable k=3**

Elbow curve:



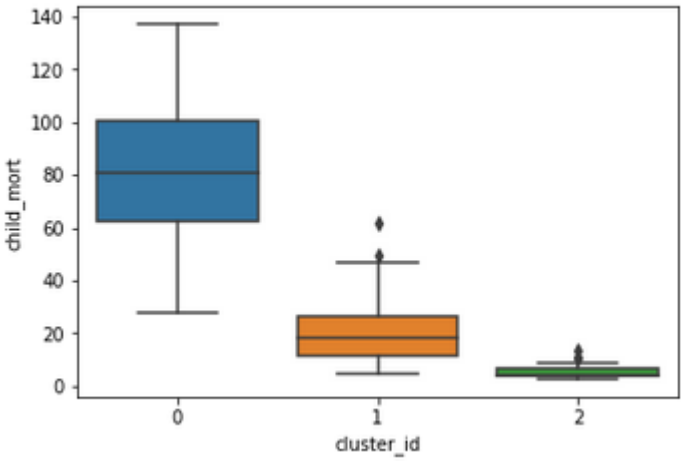
Silhouette analysis

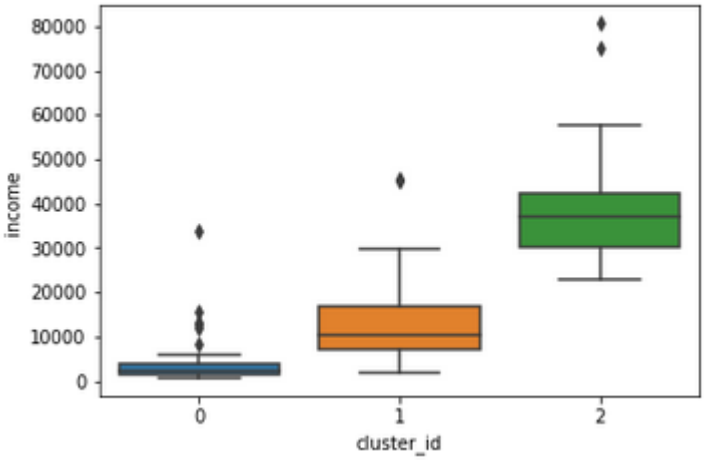
```
For n_clusters=2, the silhouette score is 0.29726057492282987
For n_clusters=3, the silhouette score is 0.30186326275710396
For n_clusters=4, the silhouette score is 0.26688062493439635
For n_clusters=5, the silhouette score is 0.25674221859745056
For n_clusters=6, the silhouette score is 0.28433267229835446
For n_clusters=7, the silhouette score is 0.28042240299010146
For n_clusters=8, the silhouette score is 0.2322489261933252
```

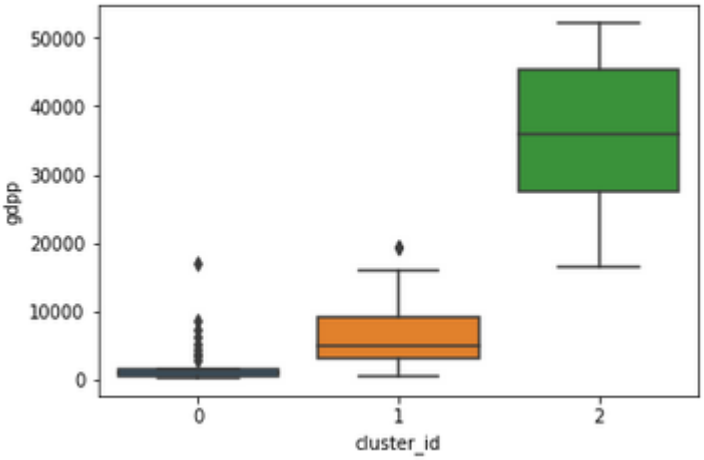
From both the elbow curve as well as the silhouette analysis, it seems k=3 seems a viable choice.

And also apart from above statistical significance, it also makes sense to choose 3 clusters in business sense as well as countries can then be divided into underdeveloped, developing and developed countries

Once the clusters were formed, the labels were assigned to the original dataset and variables such as income, gdpp and child_mort were analysied against these clusters to come identify what clusters were formed.







From the above three plots, we can clearly see that countries are divided into three clusters of Developed (cluster 2), Developing (cluster 1) and Under-developed (cluster 0) countries wherein the child mortality rate is highest in under-developed countries or cluster 0 whereas income and gdpp of these countries is lowest.

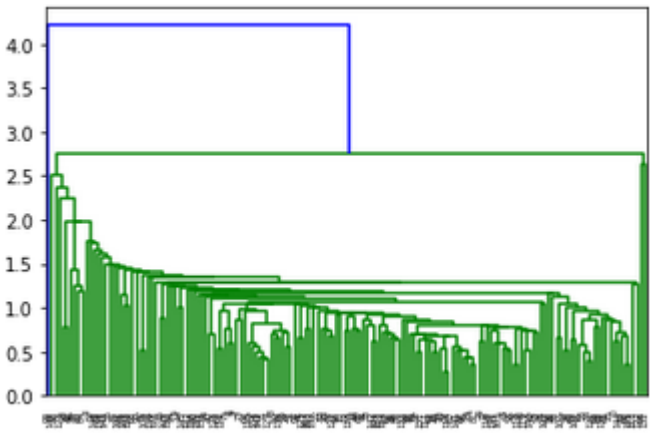
Similarly, we can see that child mortality for developed countries or cluster 2 is lowest and income and gdpp of these countries is highest.

For developing countries or cluster 1, all these variables are somewhere in between of the above 2

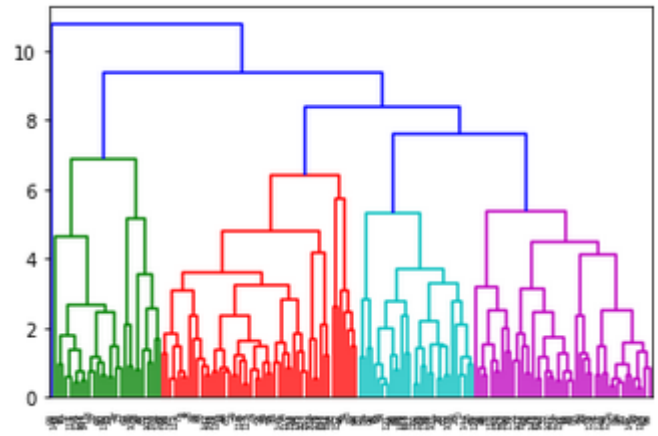
Hierarchical clustering

For hierarchical clustering, I checked with both single and complete linkages but chose complete linkage to go for further analysis as the dendrogam with complete linkage was much more interpretable as can be seen below:

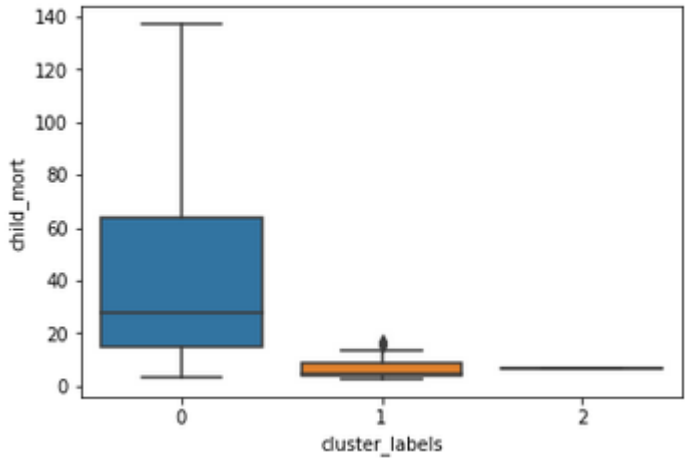
Single linkage

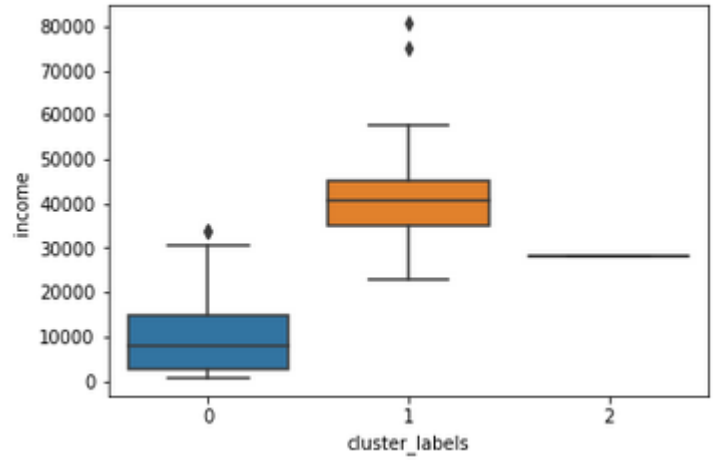


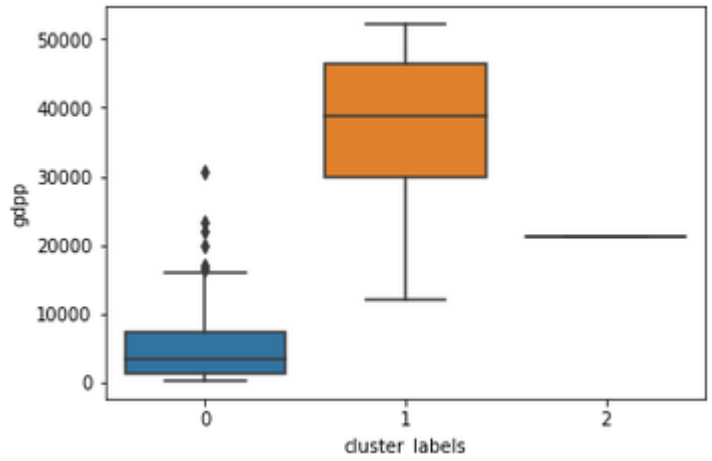
Complete linkage



After choosing complete linkage for further analysis, I cut the tree to form three clusters and then again analysed the clusters again variables such as income, gdpp and child_mort.







From the above 3 clusters formed by Hierarchical clustering, we can observe that there cluster 2 only has one country associated with it and hence doesn't really make any sense. However cluster 0 and cluster 1 form an identifiable pattern where cluster 0 seems to be formed of under developed countries wherein the child mortality rate is highest and income and gdpp are lowest.

on the contrary, cluster 1 seems to be of developed countries wherein child mortality rate is lowest and income and gdpp is highest. hence obviously, countries from cluster 0 are in real need of aid.

Comparing K-means and Hierarchical clustering results

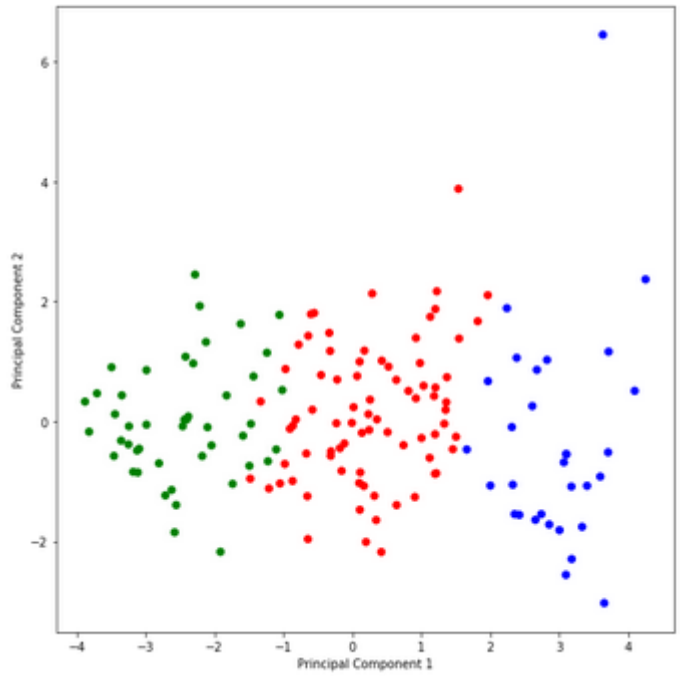
Looking at the cluster formed by K-means and Hierarchical clustering algorithm above, we can see that clusters formed by K-means are uniform with good number of countries falling under each cluster formed. However same was not the case with Hierarchical clustering wherein one cluster formed had only one or two country associated with it.

Therefore I believe K-means clustering is a better of choice of algorithm for this problem at hand.

And hence to find out the countries in direst need of aid, I got the underdeveloped countries found in cluster 0 of k-means algorithm into a separate dataframe and then sorted them to find countries with highest child mortality and lowest gdpp and income and finally came up with the below list of countries that need the aid the most:

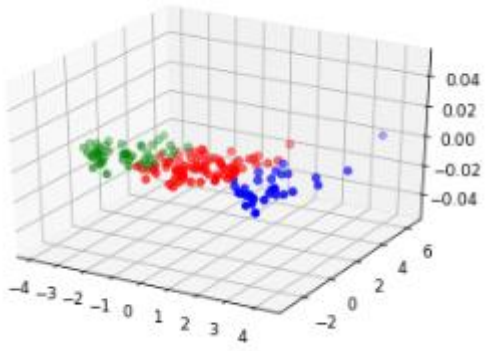
- **Congo, Dem. Rep.**
- **Niger**
- **Burundi**
- **Mozambique**
- **Togo**
- **Madagascar**
- **Guinea**
- **Guinea-Bissaue**

Visualizing the k-means clusters with principal components X-Y axes

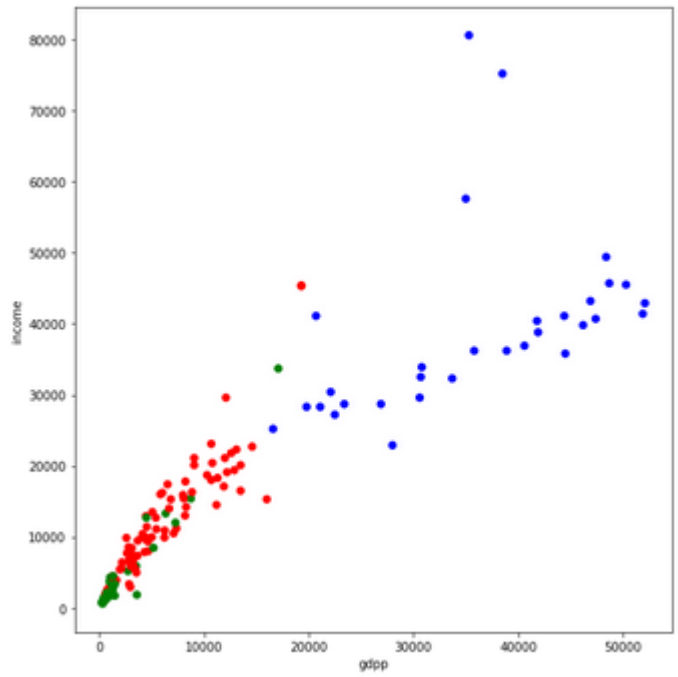


From above, we can definitely see a pattern between the clusters. Individual countries in each clusters are closely grouped together and show an increasing trend on PC1 as the cluster moves from underdeveloped to developed countries

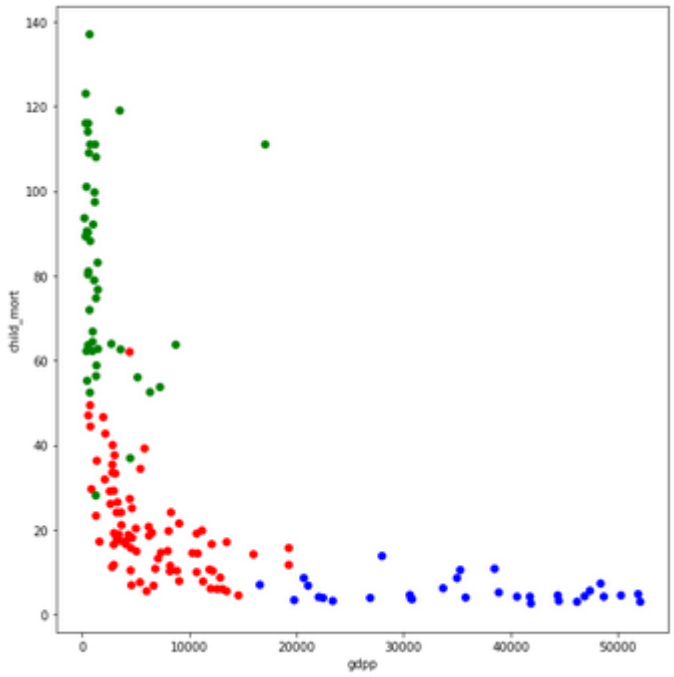
3-D



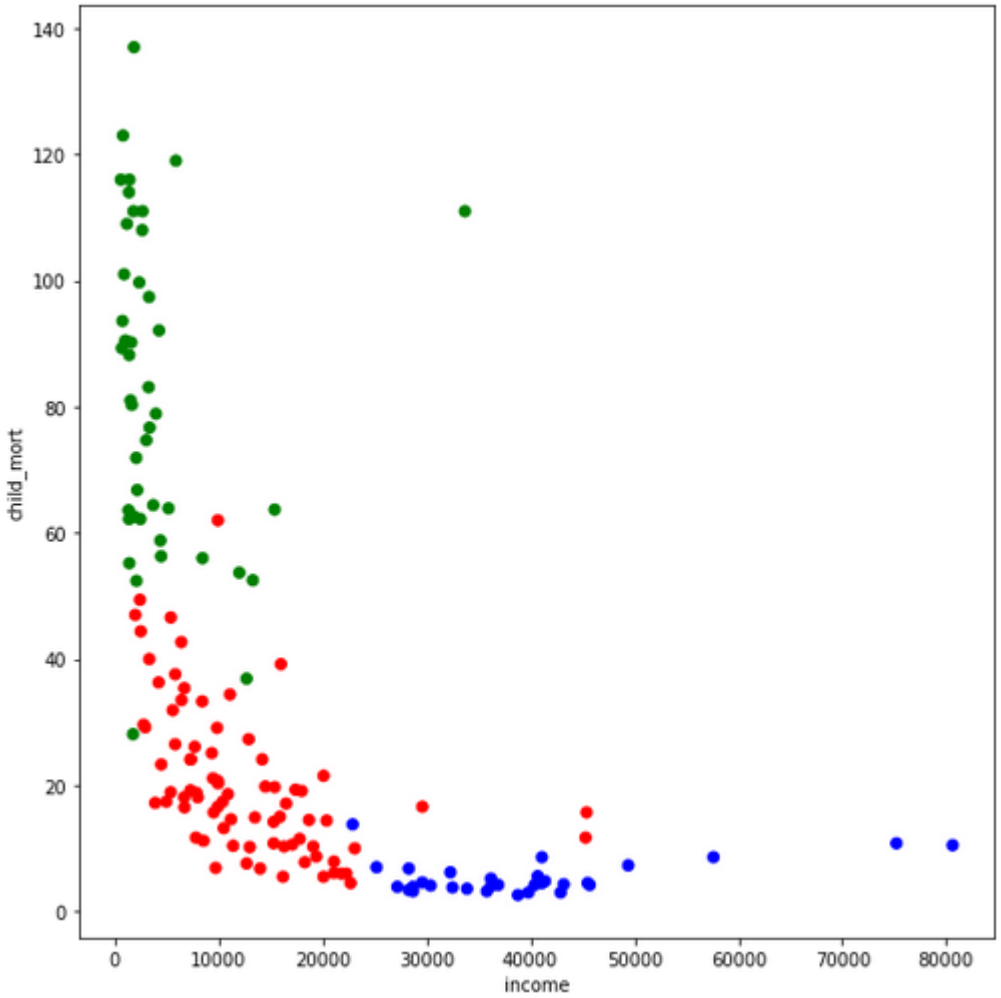
Visualizing the k-means clusters with original variables on X-Y axes



Plotting the clusters against income and gdpp, we can see that as the clusters move from underdeveloped(green) to developed(blue), both the income as well as the gdpp of the countries increases.



Plotting the clusters against child mortality and gdpp, we can see that as the clusters move from underdeveloped(green) to developed(blue), the child mortality decreases whereas the gdpp increases.



Same pattern can be seen while plotting the clusters against child mortality and income wherein as the clusters move from underdeveloped(green) to developed(blue), the child mortality decreases whereas the income increases.