CLUSTERING AND PCA

Problem statement and analysis approach

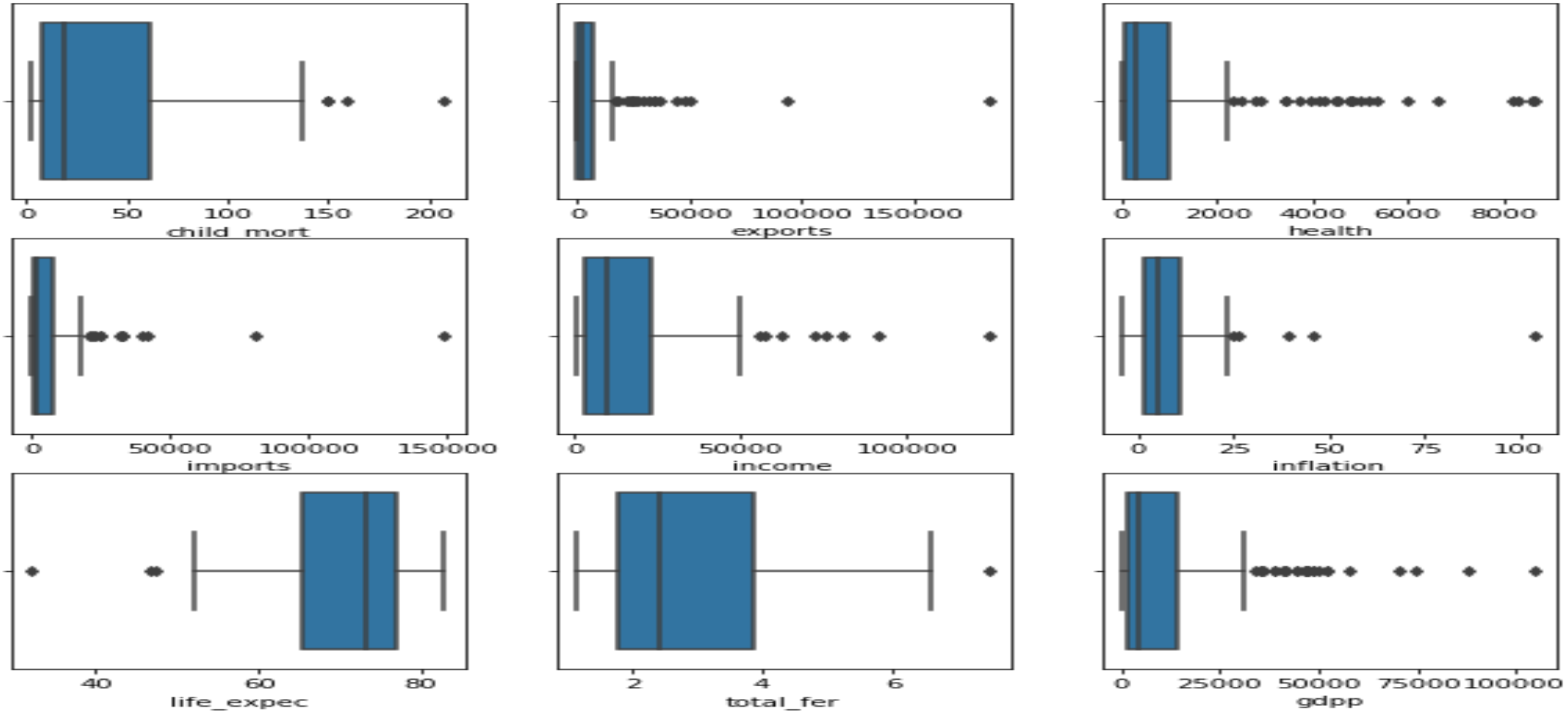
In ‘Clustering of Countries’ assignment the problem statement is that HELP is an NGO which was fighting against poverty by providing the basic amenities to the countries which are effected due to natural calamities. The CEO of NGO wants to select the countries that are highly required the help.

For this the methodology used was ‘Clustering’ for selecting the countries that highly required the help. Two types of clustering algorithms are used: 1. K-Means, 2. Hierarchical clustering. And also PCA method is used to reduce the dimensionality

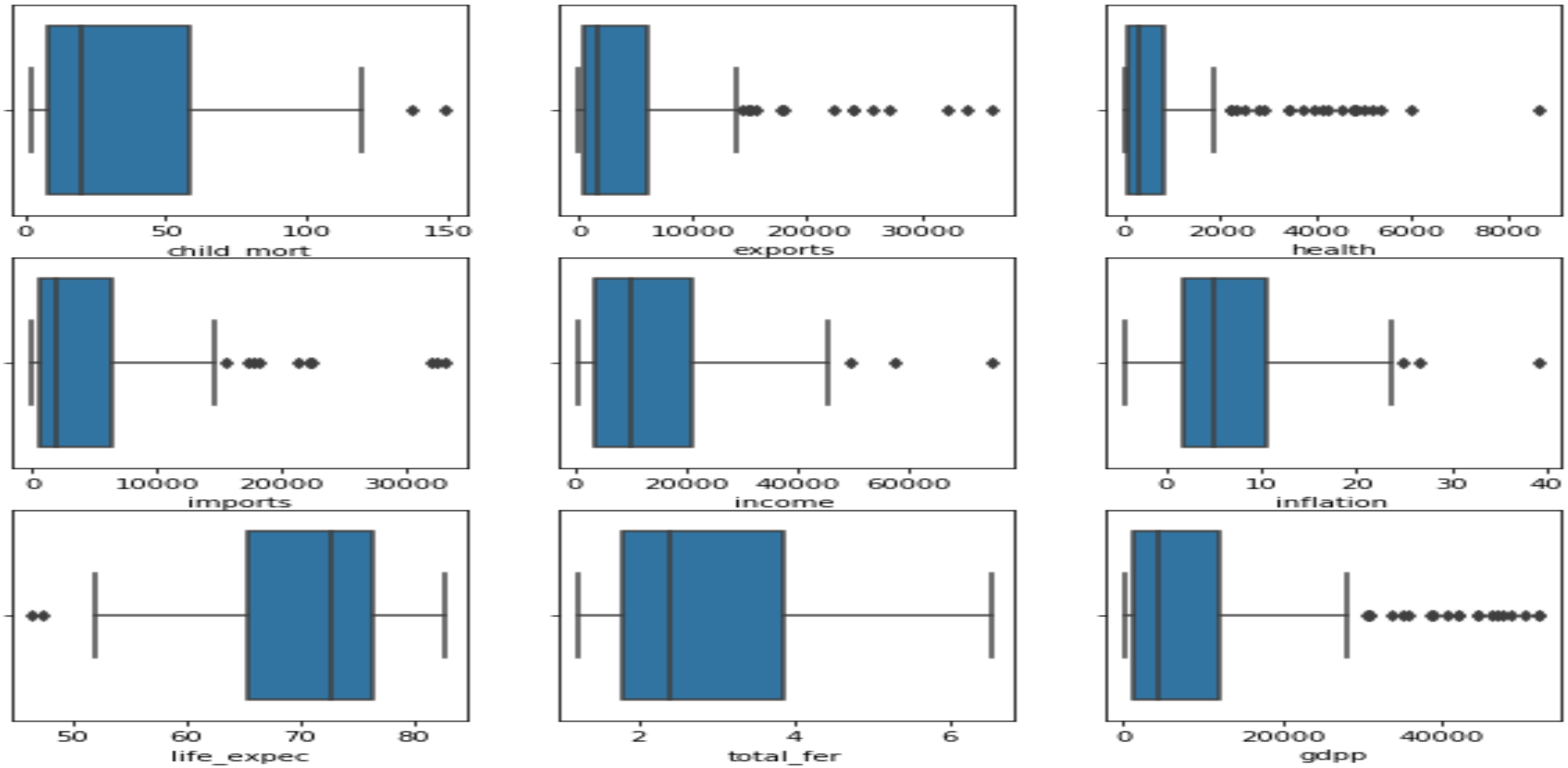
Analysis part

- First we check the shape which have (167,10) rows and columns.
- The columns like ‘exports’, ’imports’, ‘health’ are in %gdp so, converted them to actual values.
- After that we check for the outliers in the dataframe and remove some of them as removing number of outliers will have loss of information. So, small amount of outliers are removed.
- The final dataframe after performing outliers have 153 rows.

Without outlier removal



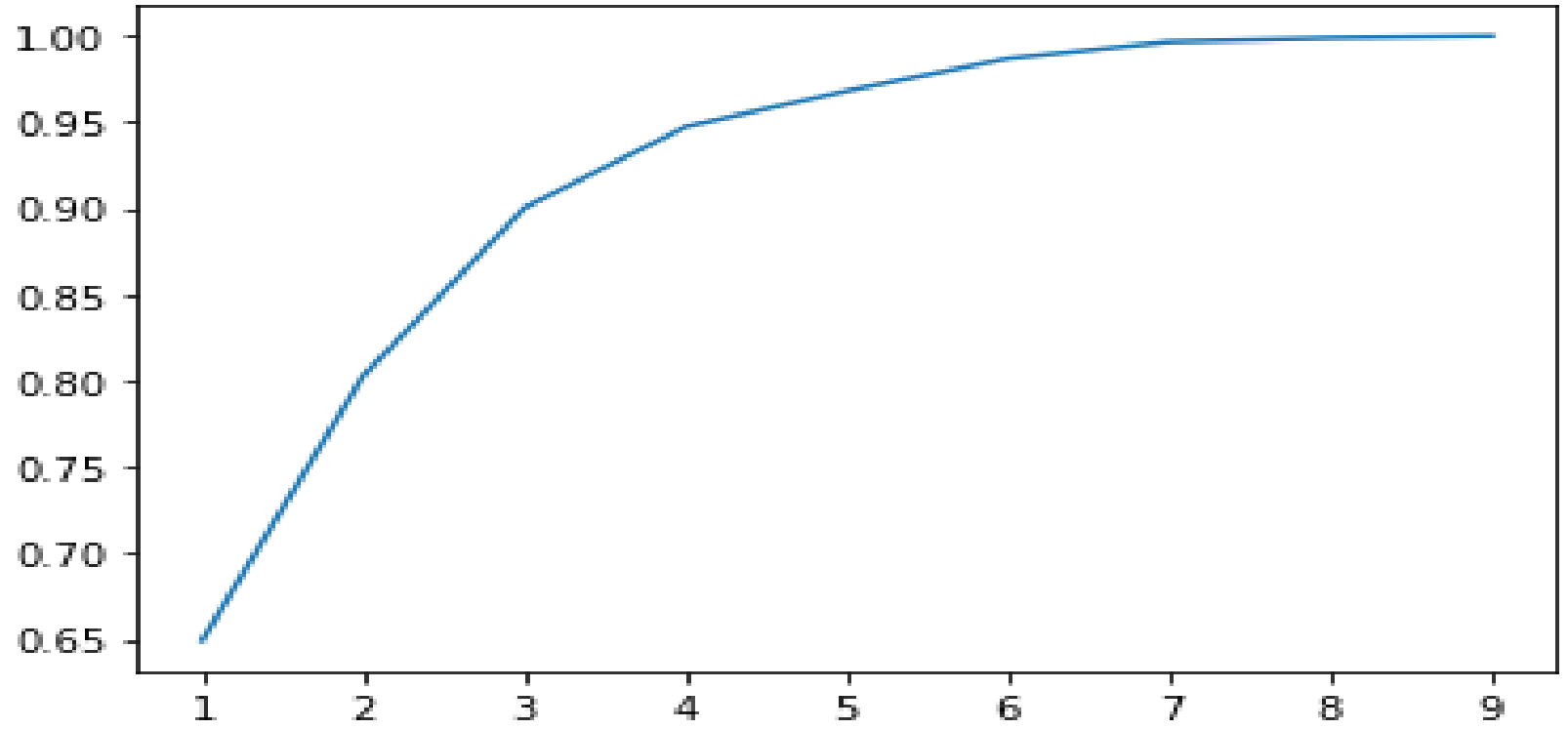
With removing outliers



Principal Component Analysis

- Actually for performing the clustering methods with original dataframe we are using principal component analysis (PCA) method which is used to reduce the dimensionality.
- In PCA method we will get the principal components by those we will perform the clustering.
- In PCA first we will scale, fit and transform the data.
- By finding the variance of each column the income is having highest variance and after that gdpp is having high.
- The first component is having 65% of variance.

By using scree plot we will select the number of principal components.



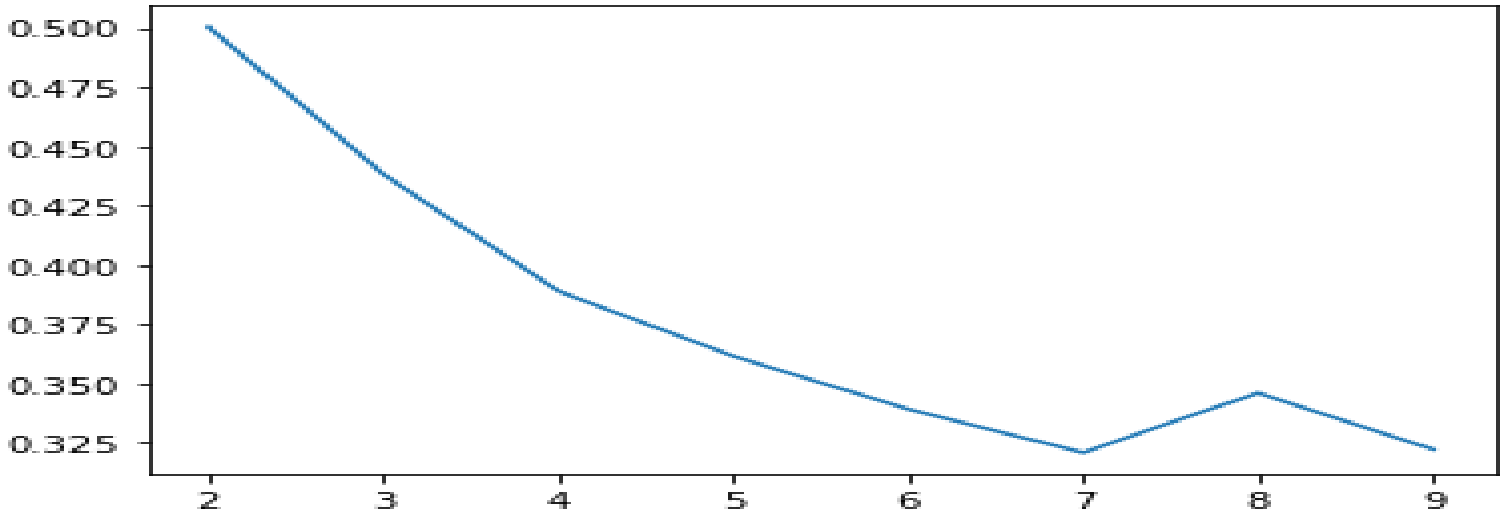
Number of PC'S we can go with are 4 which gives 95% of variance approx..

- Finally we will perform the dimensionality reduction with 4 PC components
- With final dataframe we will fit and transform the data.
- We use Hopkins measure to check how much percentage is our approach is correct. For final datafrmae the Hopkin score is 87%.

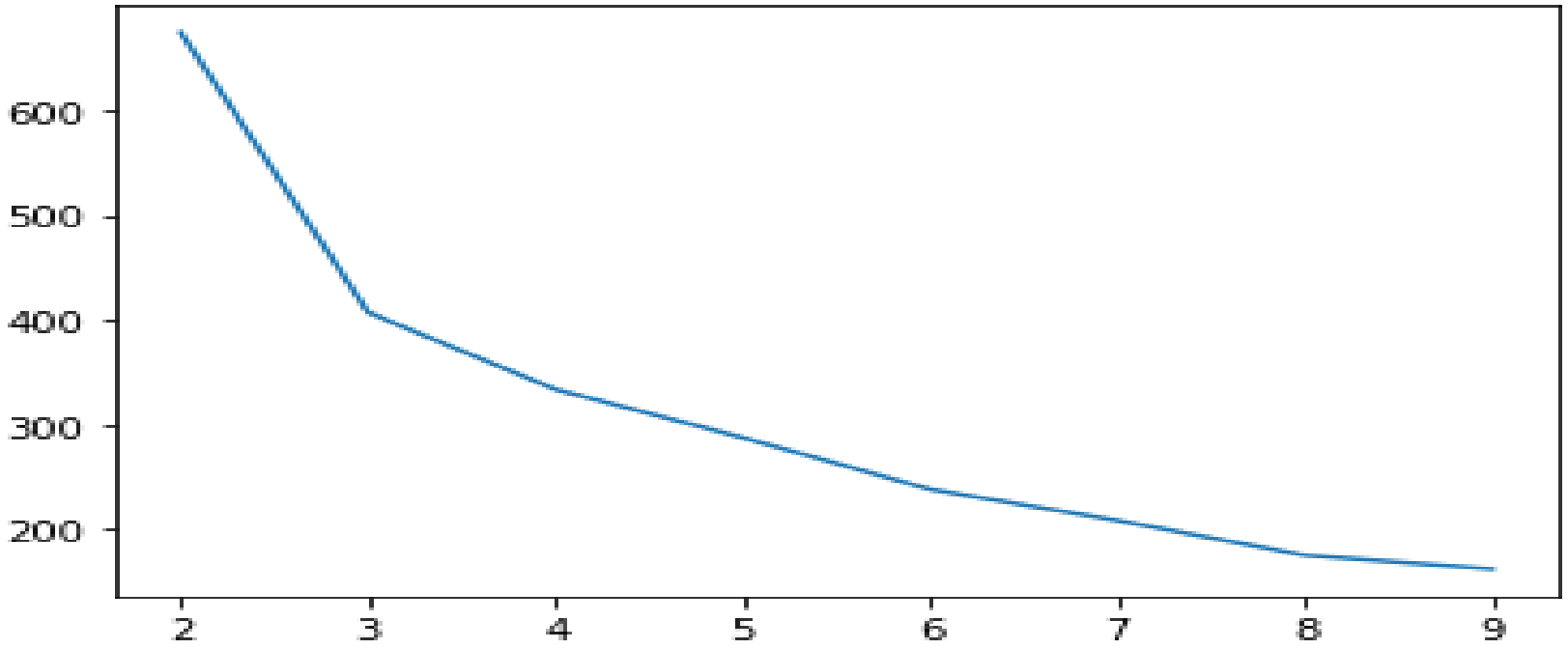
K-Means Clustering

- For selecting number of clusters we are using 2 methods.

1. silhouette score plot

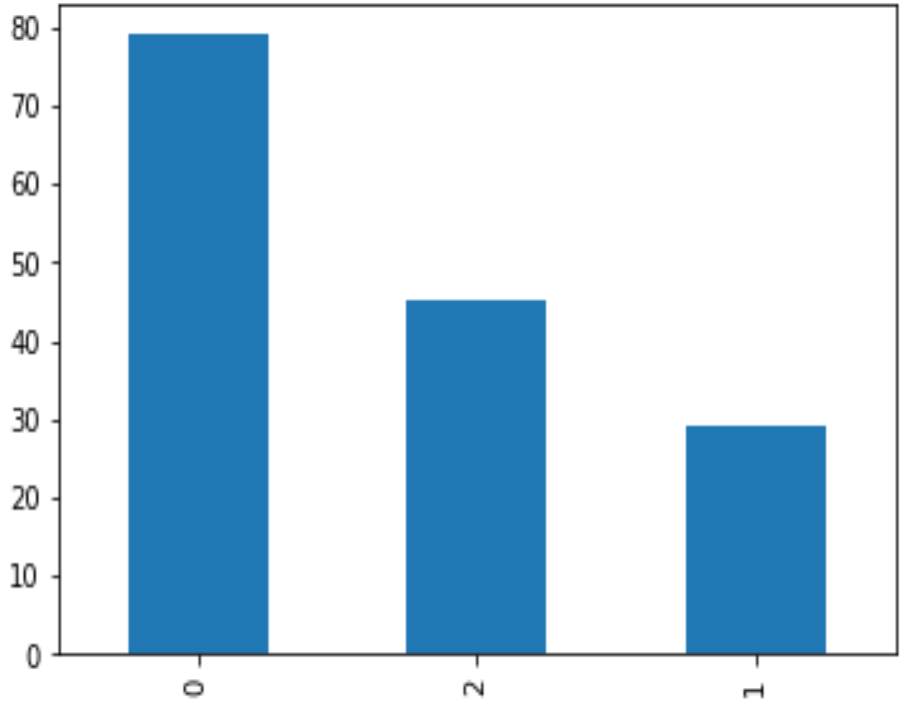


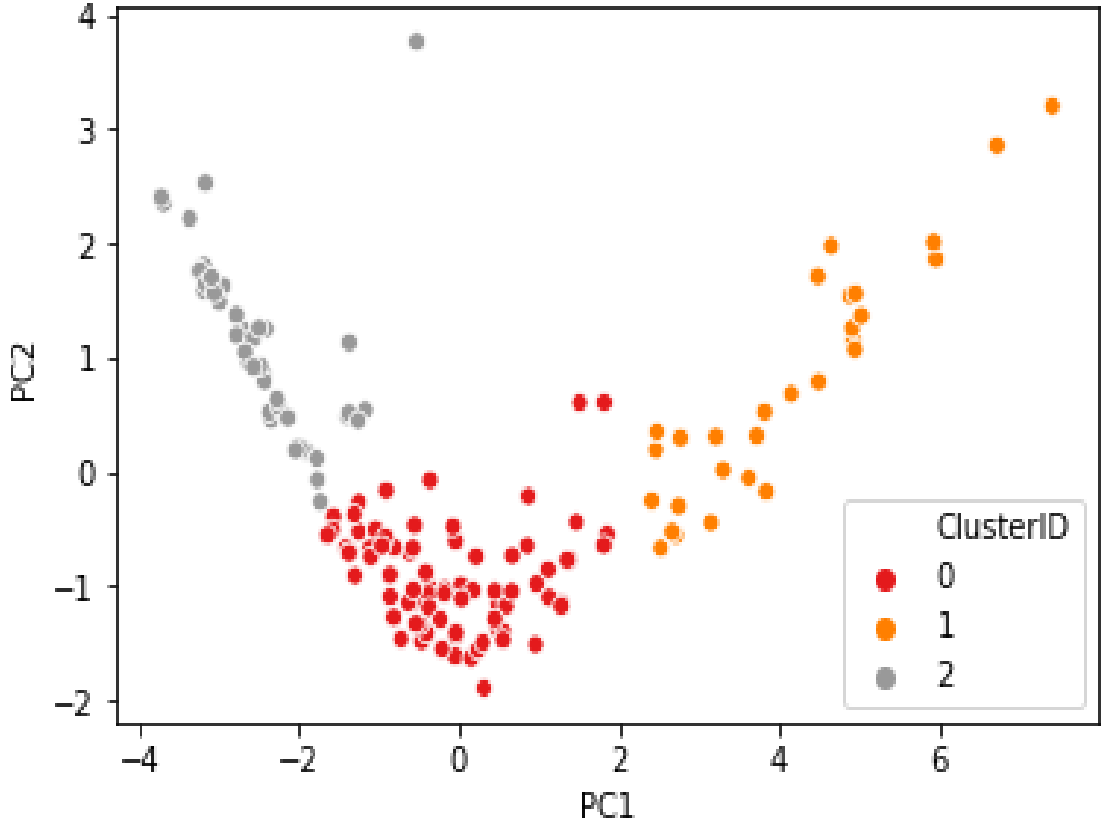
2. Elbow curve method



Form both the graphs we are selecting the number of clusters to be 3.

The datapoints that are assigned to the clusters are given by the below images

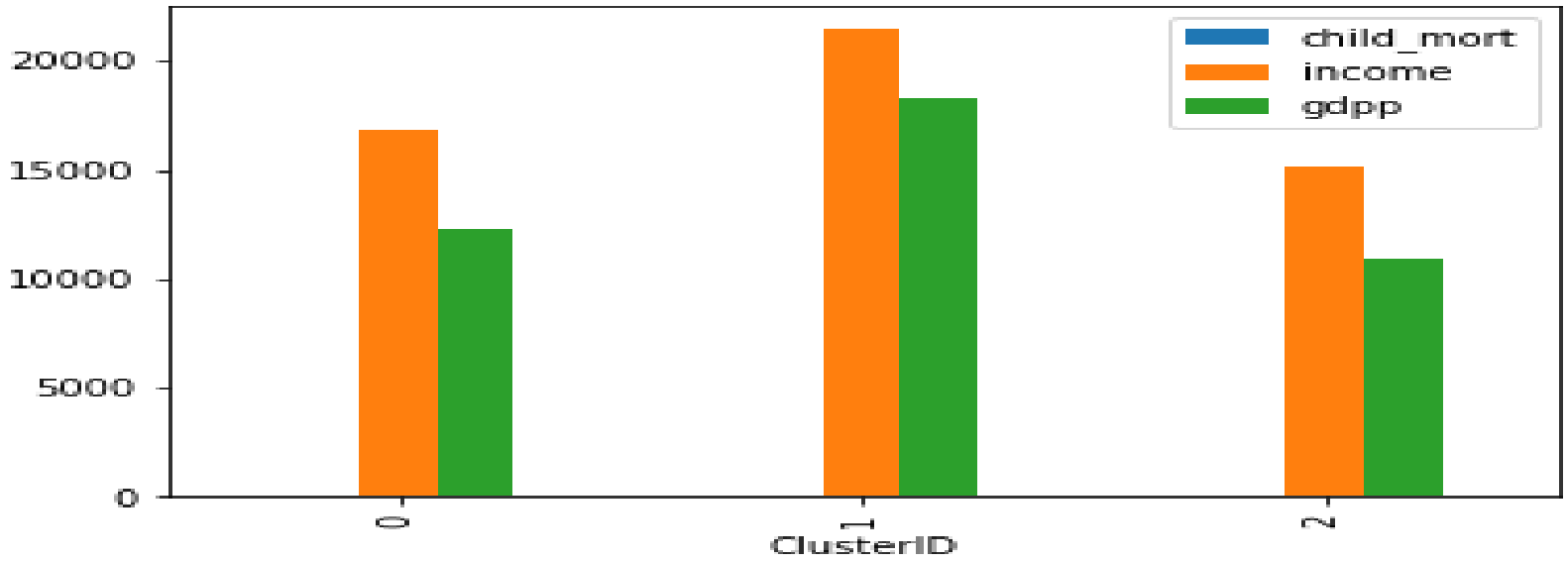




The cluster 0 is having high datapoints that are assigned.

Finally, for analyzing the clusters we are using the three cloumns [child_mort, income, gdpp] from the original dataset.

- In all the clusters, the income is high and gdpp is low, but child_mort is very very low as it was not visible in the graph
- But among all the clusters the cluster-1 is having high values and cluster-2 have low values

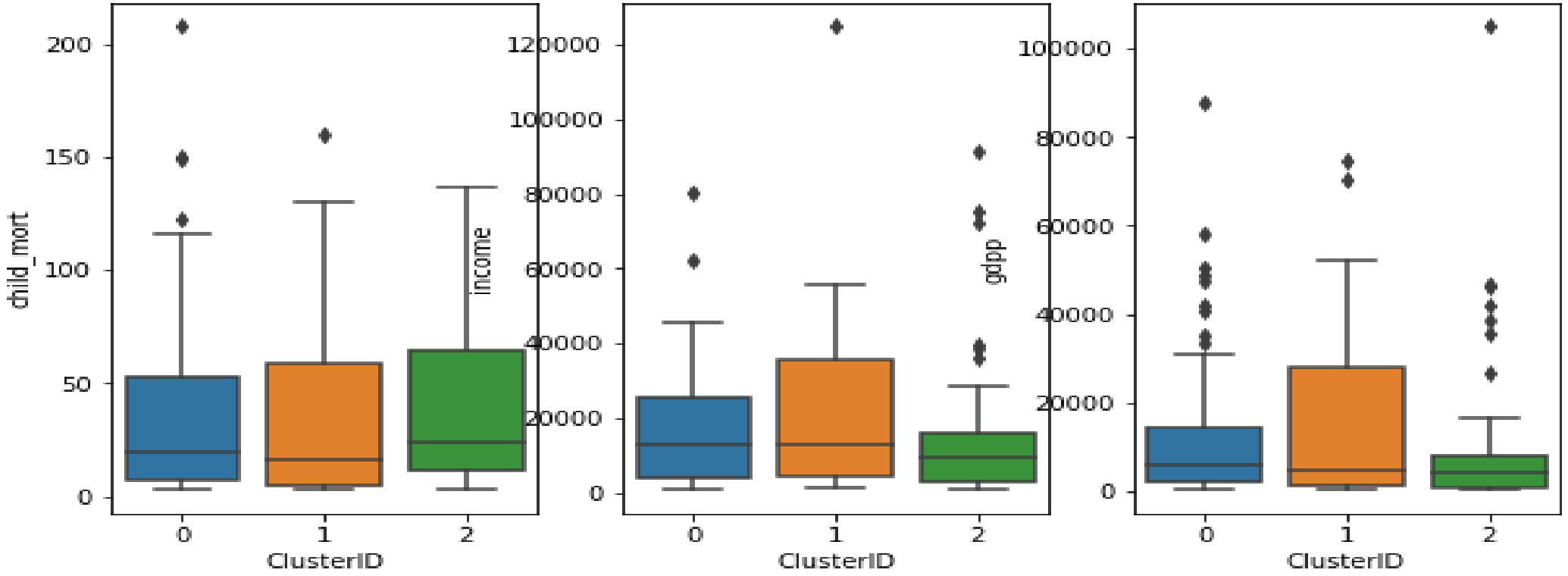


Using boxplots comparing the 3 clusters for each 3 columns

For child_mort column the clusters have somewhat minimum changes in the values, where the cluster-0 is low and cluster-2 is high.

For income column the cluster-1 have high and cluster-2 is very low.

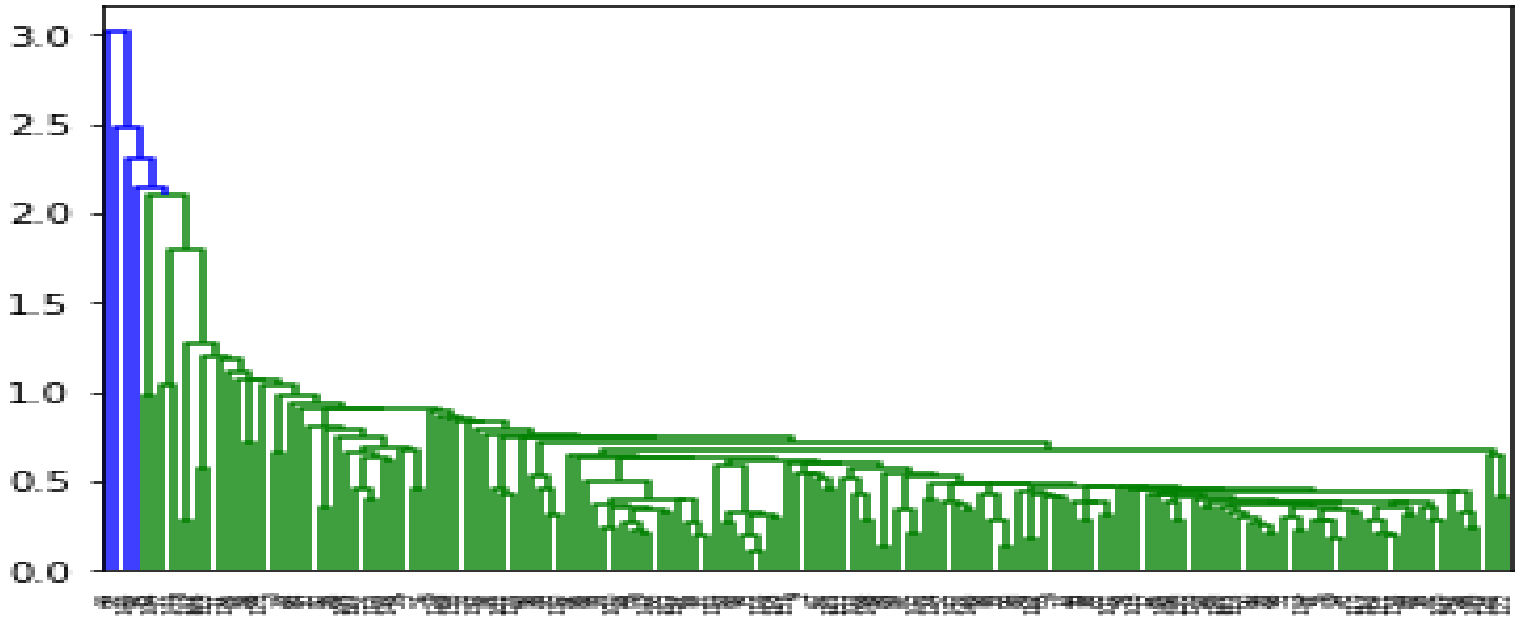
For gdpp column the cluster-1 have high and cluster-2 is very low.



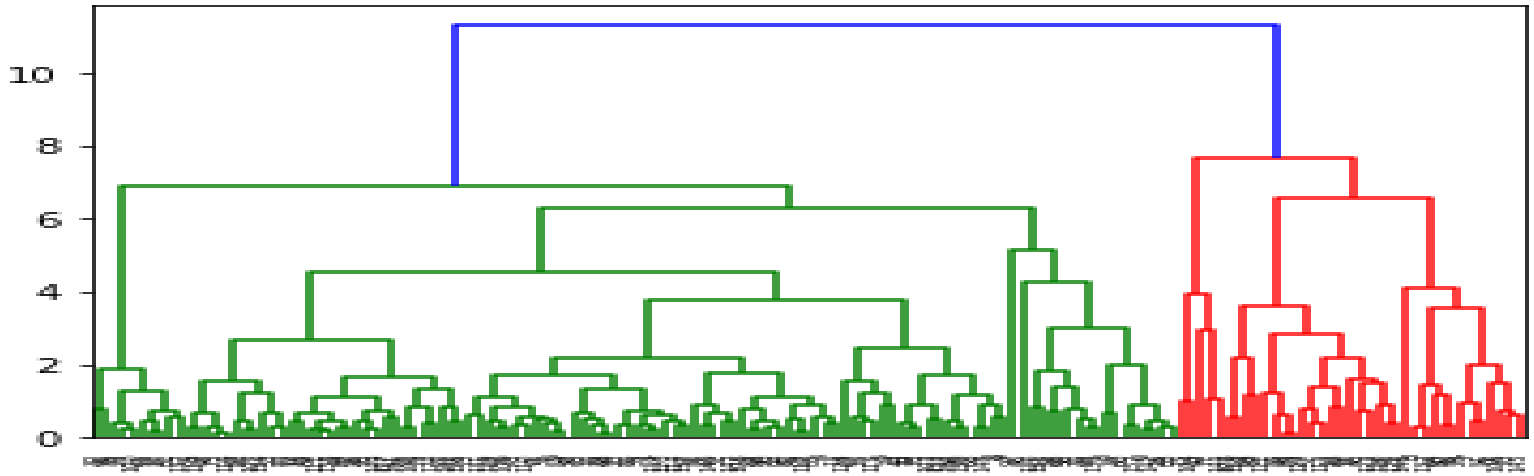
Hierarchical Clustering

In Hierarchical Clustering we are using two methods to select the number of clusters.

1. Single Linkage



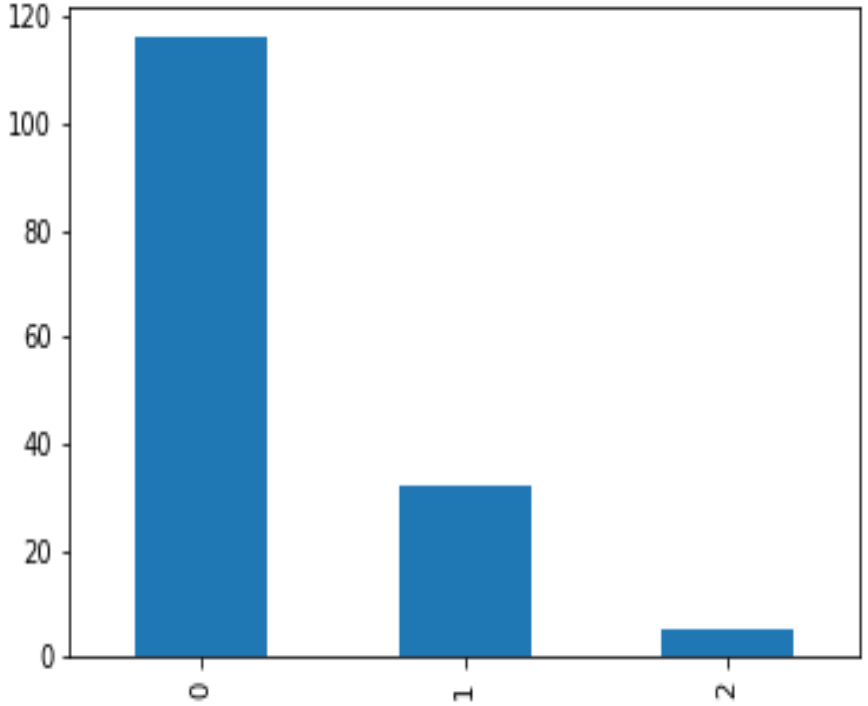
2. Complete Linkage

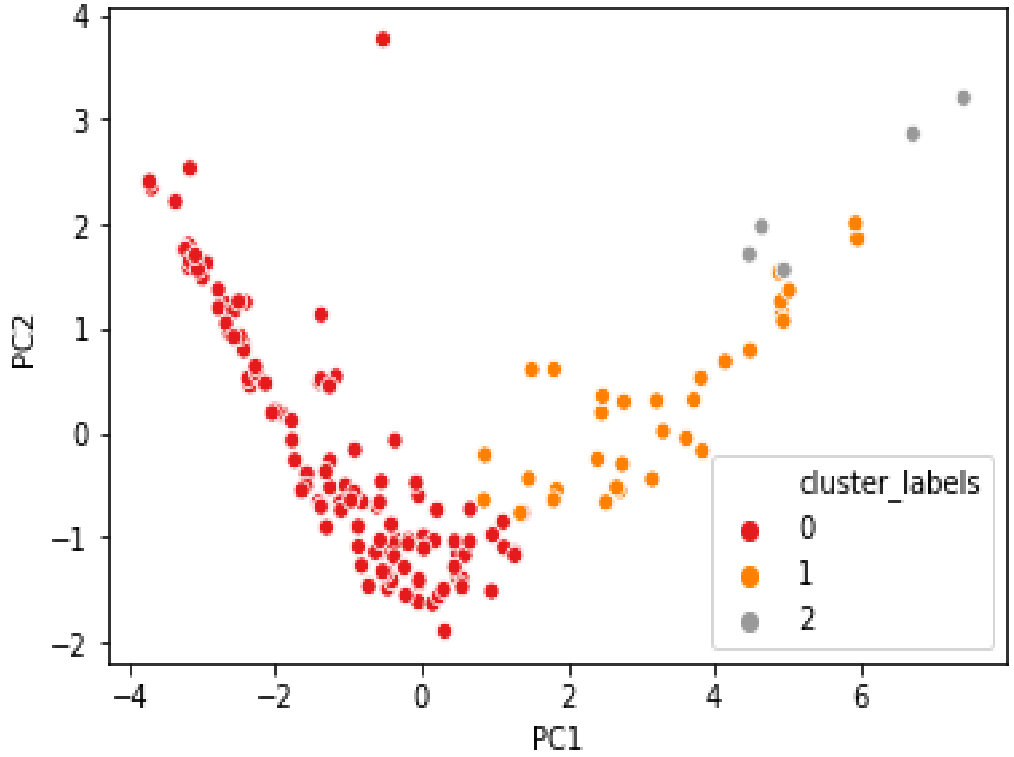


Compared to single Linkage, the complete linkage graph is having some clear view to know how clusters are formed and also we can select the clusters by viewing the graph.

So, from this we are selecting the number of clusters are 3.

The datapoints that are assigned to the clusters are given by the below images in '**Hierarchical Clustering'**

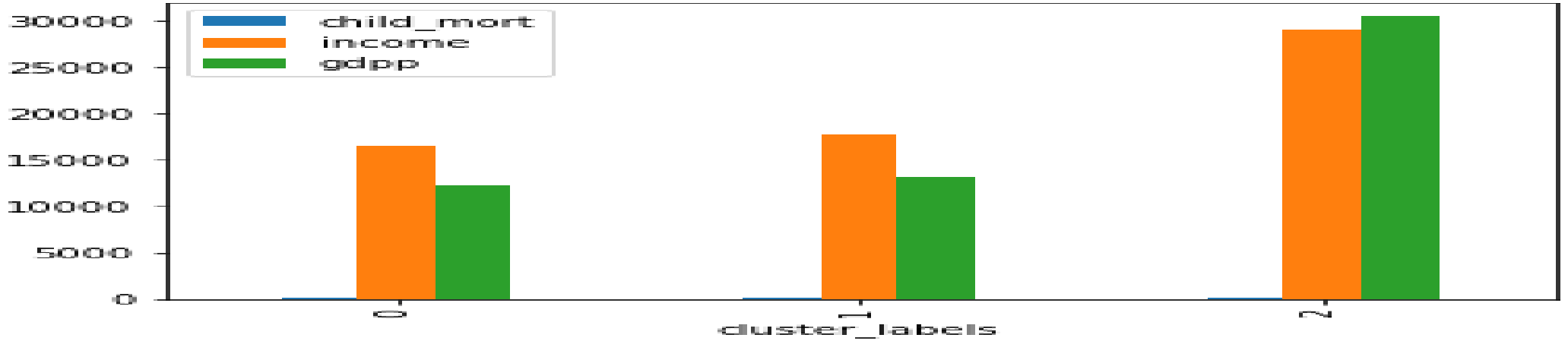




The cluster 0 is having high datapoints that are assigned

Finally, for analyzing the clusters we are using the three cloumns [child_mort, income, gdpp] from the original dataset.

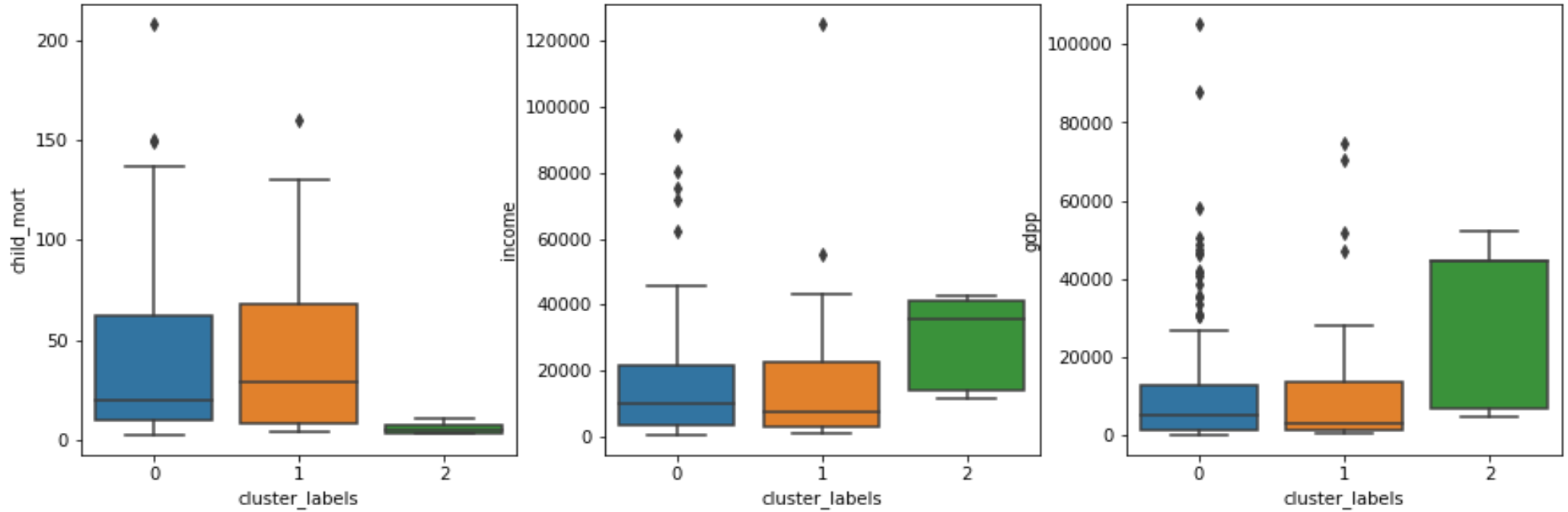
In clusters 0 & 1, the income is high and gdpp is low, but for cluster 2 data the income is low and gdpp is some what high.
But among all the clusters the cluster-2 is having high values and cluster-0 & 1 approximately are same but with slight variation the cluster-0 is having low. But here also the child_mort is very very low not visible.



For child_mort column the cluster-2 is very very low where the cluster-1 is high compared to cluster-0.

For income column the cluster-2 is very high compared to other clusters.

For gdpp column the cluster-2 is very high than others.



These are the final list of countries that highly require the need.

- Afghanistan, Angola, Benin, Botswana, Bulgaria, Burkina Faso, Cambodia, Cape Verde, China, Colombia, Comoros, Congo, Rep., Ecuador, Egypt, Fiji, Finland, Gambia, Greece, Grenada, Iceland, Italy, Jamaica, Kazakhstan, Kuwait, Kyrgyz Republic, Lesotho, Liberia, Luxembourg, Madagascar, Mali, Mauritius, Moldova, Myanmar, Pakistan, Peru, Romania, Russia, Serbia, Singapore, Slovak Republic, Solomon Islands, South Africa, Sudan, Tonga, Tunisi.