Clustering & PCA Assignment

ASHUTOSH KUMAR

Contact: ashutoshind2017@outlook.com
+91-8904866645

fppt.com

BACKGROUND

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 programs, 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.** The datasets containing those socio-economic factors and the corresponding data dictionary are provided.

fppt.com

PROBLEM STATEMENT

Objectives

Your main task is to cluster the countries by the factors mentioned above and then present your solution and recommendations to the CEO using a PPT. The following approach is suggested :

Perform PCA on the dataset and obtain the new dataset with the Principal Components. Choose the appropriate number of components k. You need to perform your clustering activity on this new dataset, i.e. the PCA modified dataset with the k components.

Outlier Analysis: You must perform the Outlier Analysis on the dataset, before or after performing PCA, as per your choice. However, you do have the flexibility of not removing the outliers if it suits the business needs or a lot of countries are getting removed. Hence, all you need to do is find the outliers in the dataset, and then choose whether to keep them or remove them depending on the results you get.

Try both K-means and Hierarchical clustering(both single and complete linkage) on this dataset to create the clusters. [Note that both the methods may not produce identical results and you might have to choose one of them for the final list of countries.]

Analyse the clusters and identify the ones which are in dire need of aid. You can analyse the clusters by comparing how these three variables - [**gdpp**, **child_mort** and **income**] vary for each cluster of countries to recognise and differentiate the clusters of developed countries from the clusters of under-developed countries. Note that you perform clustering on the PCA modified dataset and the clusters that are formed are being analysed now using the original variables to identify the countries which you finally want to select.

Also, you need to perform visualisations on the clusters that have been formed. You can do this by choosing the first two Principal Components (on the X-Y axes) and plotting a scatter plot of all the countries and differentiating the clusters. You should also do the same visualisation using any two of the original variables (like gdpp, child_mort, etc.) on the X-Y axes as well. You can also choose other types of plots like boxplots, etc.

The final list of countries depends on the number of components that you choose and the number of clusters that you finally form. Also, both K-means and Hierarchical may give different results. Hence, there might be some subjectivity in the final number of countries that you think should be reported back to the CEO. Here, **make sure that you report back at least 5 countries which are in direst need of aid from the analysis work that you perform.**

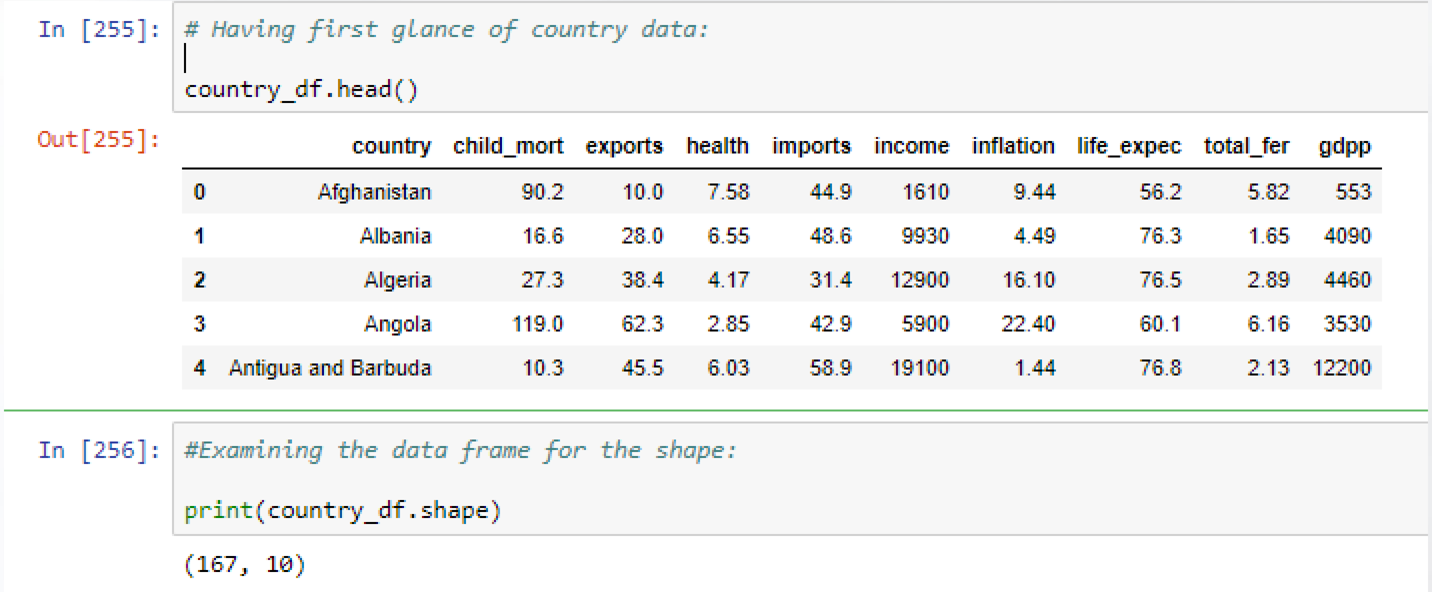
UNDERSTANDING THE DATA

Data Dictionary for the data:

Column Name	Description
country	Name of the country
child_mort	Death of children under 5 years of age per 1000 live births
exports	Exports of goods and services. Given as %age of the Total GDP
health	Total health spending as %age of Total GDP
imports	Imports of goods and services. Given as %age of the Total GDP
Income	Net income per person
Inflation	The measurement of the annual growth rate of the Total GDP
life_expec	The average number of years a new born child would live if the current mortality patterns are to remain the same
total_fer	The number of children that would be born to each woman if the current age-fertility rates remain the same.
gdpp	The GDP per capita. Calculated as the Total GDP divided by the total population.

fppt.com

READING AND FIRST GLANCE OF DATA



```
In [255]: # Having first glance of country data:

          country_df.head()
```

Out[255]:

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
0	Afghanistan	90.2	10.0	7.58	44.9	1610	9.44	56.2	5.82	553
1	Albania	16.6	28.0	6.55	48.6	9930	4.49	76.3	1.65	4090
2	Algeria	27.3	38.4	4.17	31.4	12900	16.10	76.5	2.89	4460
3	Angola	119.0	62.3	2.85	42.9	5900	22.40	60.1	6.16	3530
4	Antigua and Barbuda	10.3	45.5	6.03	58.9	19100	1.44	76.8	2.13	12200

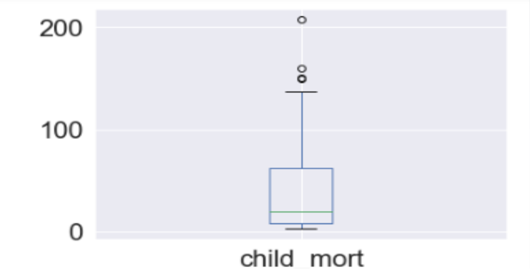
```
In [256]: #Examining the data frame for the shape:

          print(country_df.shape)

          (167, 10)
```

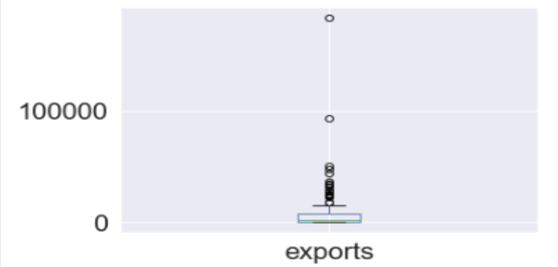
So there are 167 countries with total 10 columns for each country.

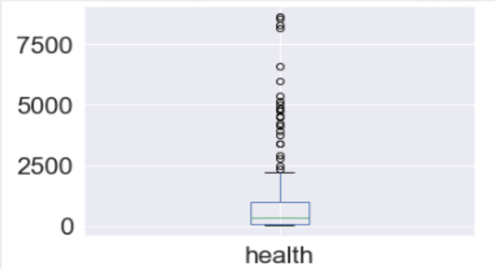
fppt.com

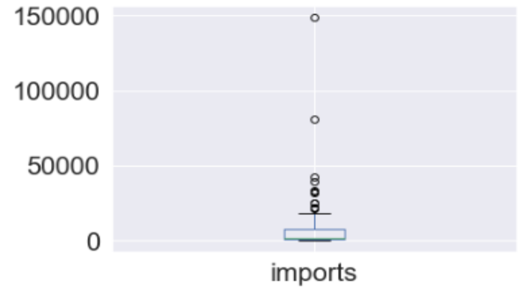
OUTLIER DETECTION IN THE DATASET

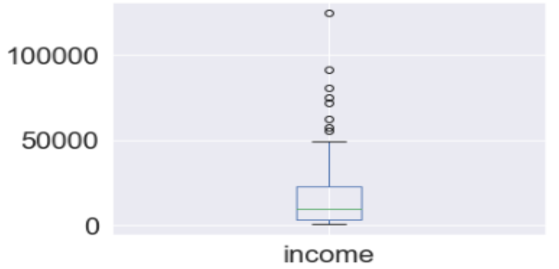
- After performing the EDA , data cleaning, feature engineering etc., next step is outlier detection.

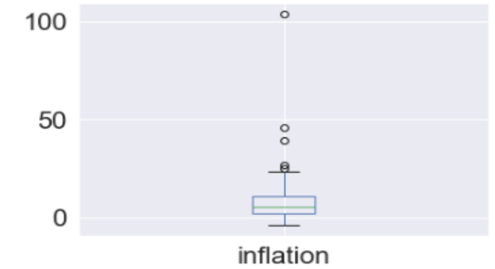


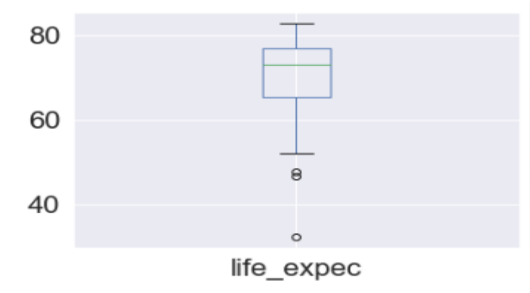


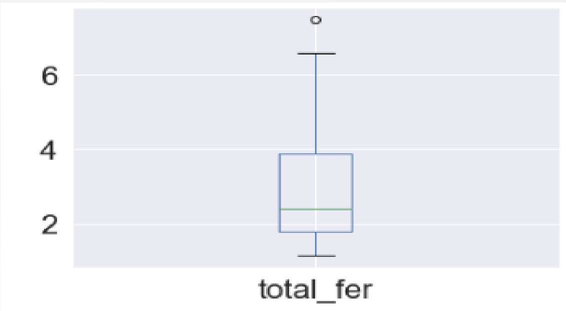


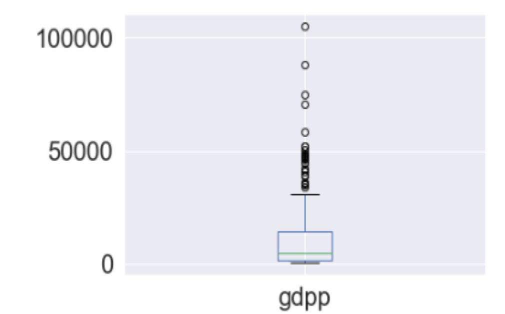






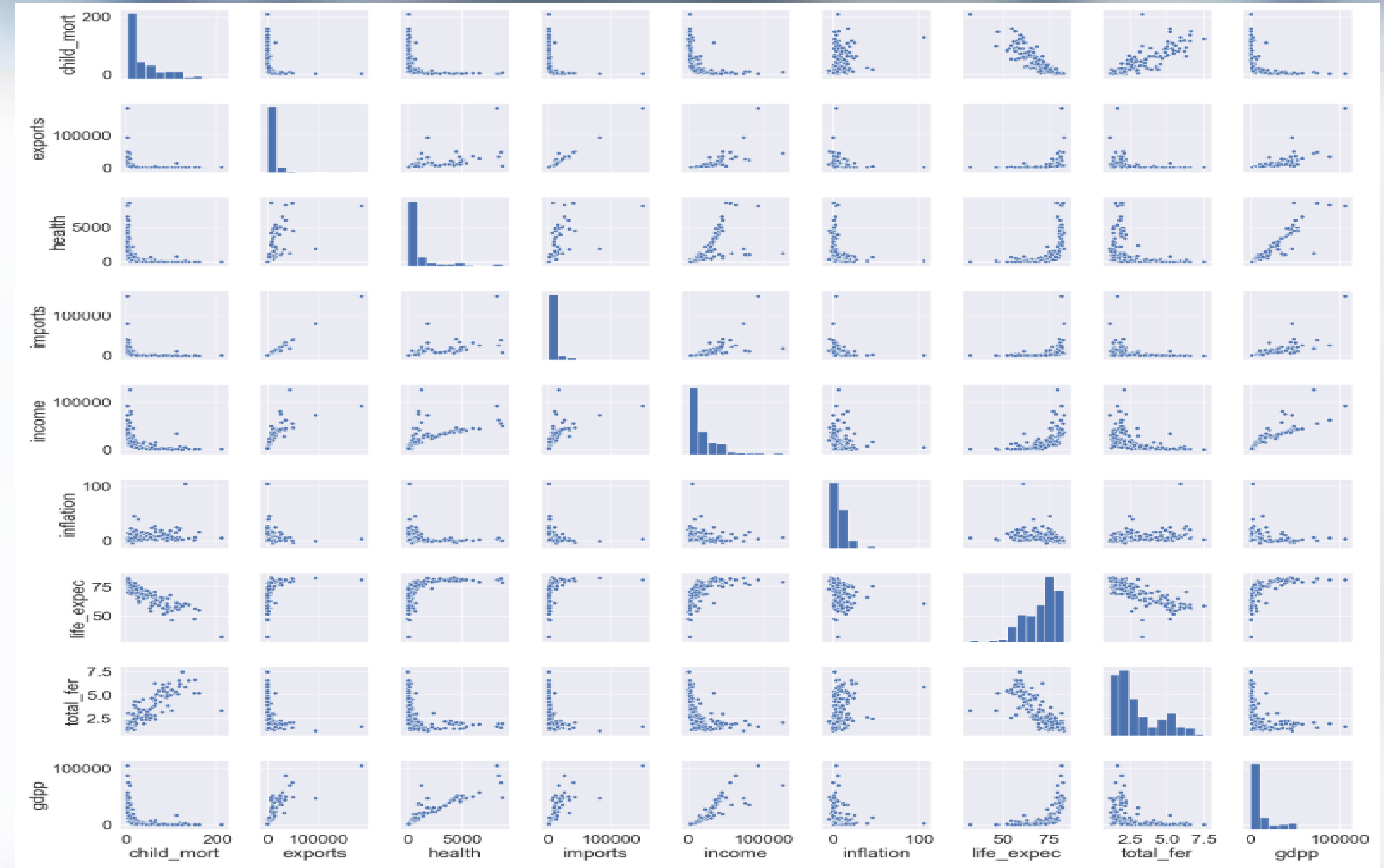




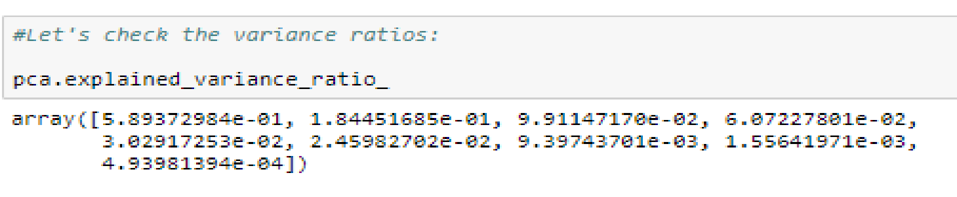


Inference: As we can see from above boxplots of numerical columns, there are few outliers namely for the columns child_mort, exports, health, imports , income, inflation, life_expec, gdpp etc.We will deal with the same after performing PCA on the same if needed.

Visualizing all the numerical variables using the pairplot

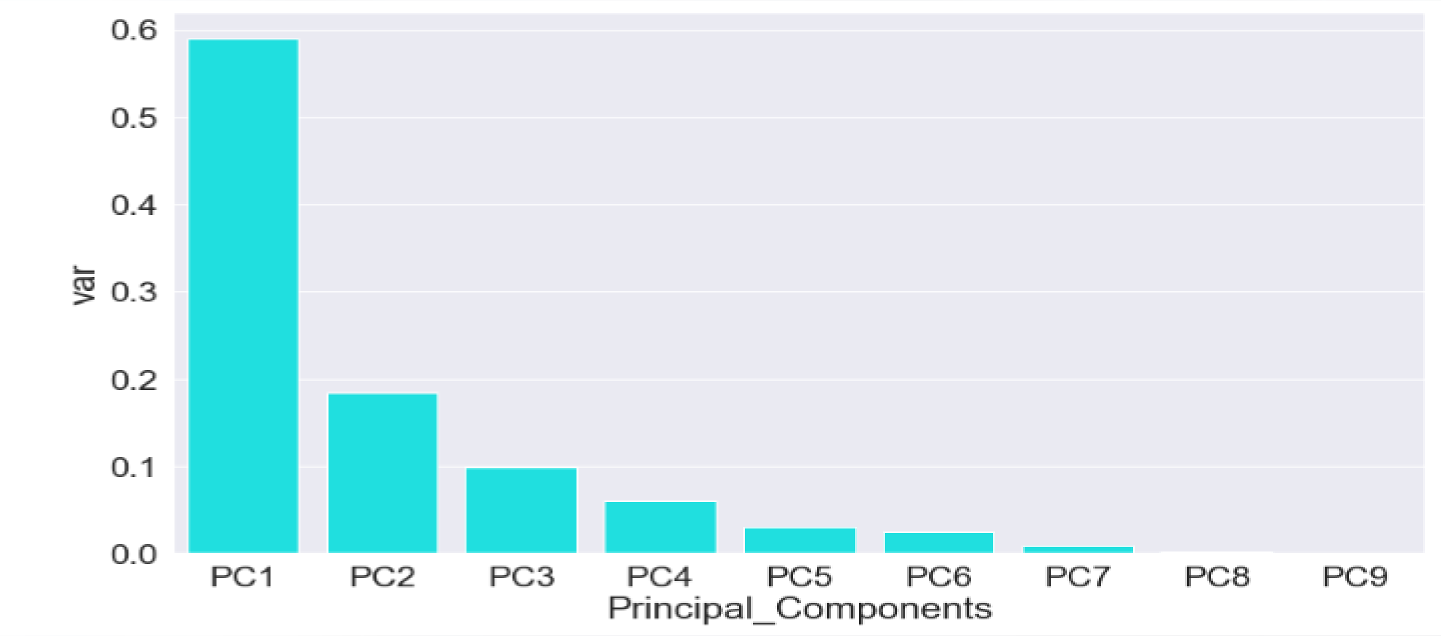


Explained Variance Ratio After PCA is applied



```
#Let's check the variance ratios:

pca.explained_variance_ratio_
```

```
array([5.89372984e-01, 1.84451685e-01, 9.91147170e-02, 6.07227801e-02,
       3.02917253e-02, 2.45982702e-02, 9.39743701e-03, 1.55641971e-03,
       4.93981394e-04])
```



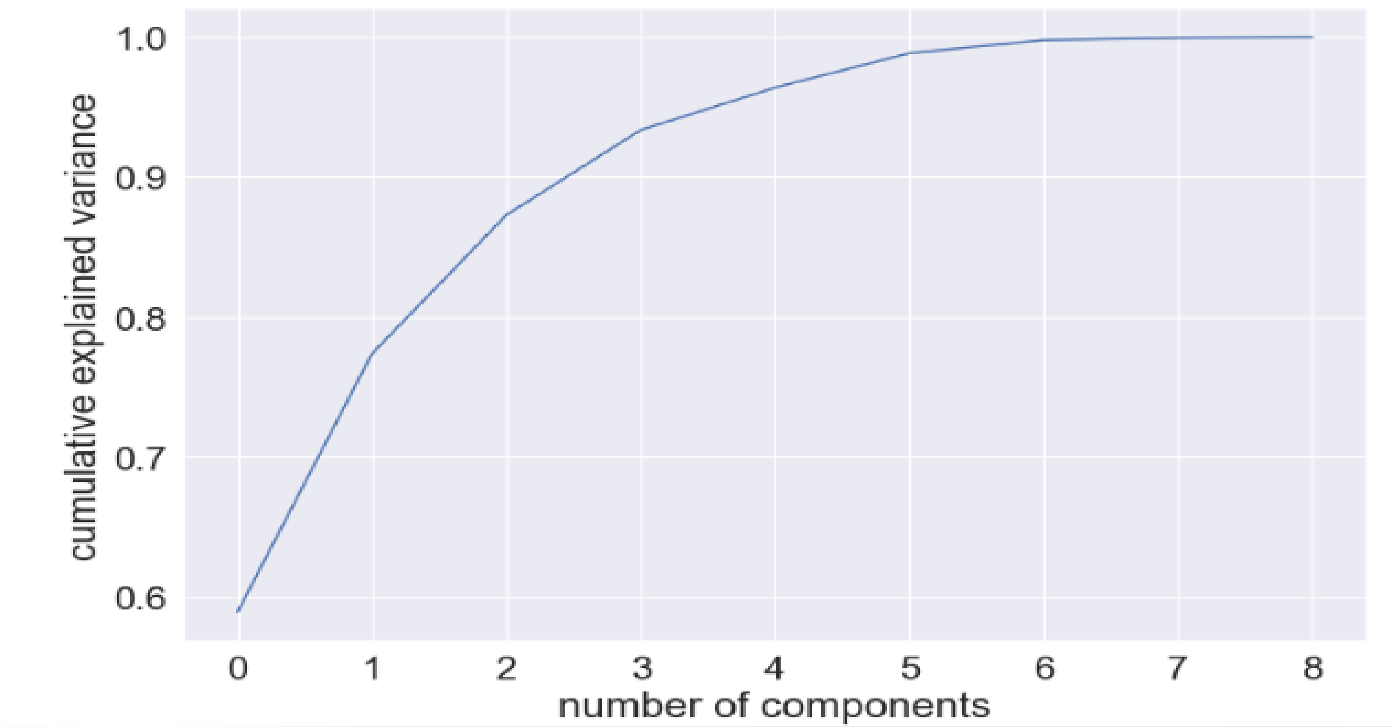
The explained variance tells how much information (variance) can be attributed to each of the principal components. This is important as while you can convert n dimensional space to lesser dimensional space, you lose some of the variance (information) when you do this. By using the attribute explained_variance_ratio_, we can see that the 1st principal component contains 58.89% of the variance,2nd principal component contains 18.44% of the variance, 3rd 9.91% and 4th 6.07% of the variance.

Inference:

Together, the four components contain around 94% of the information.

fppt.com

Scree plot for Principal Component Analysis

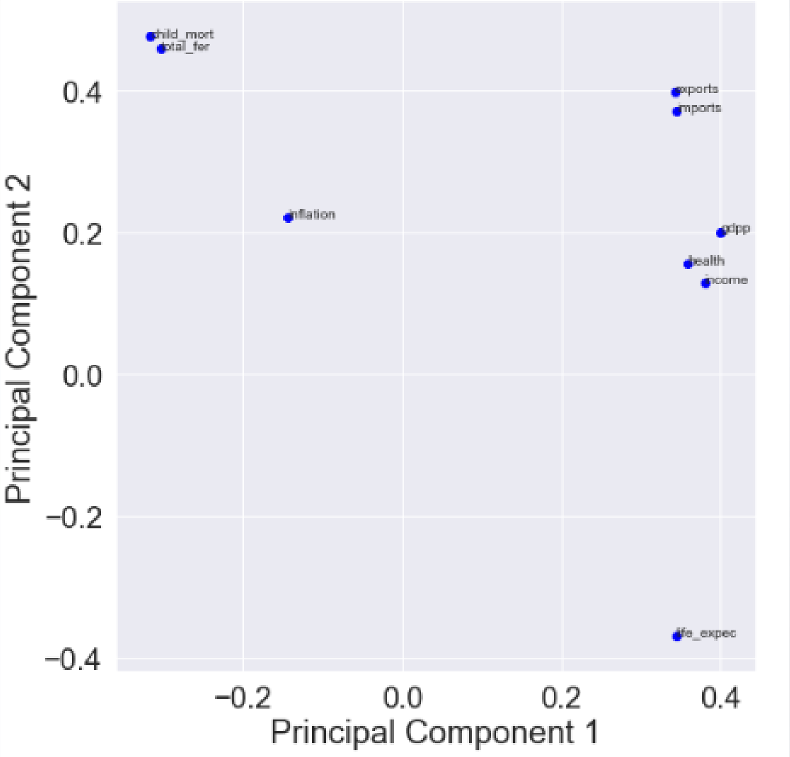


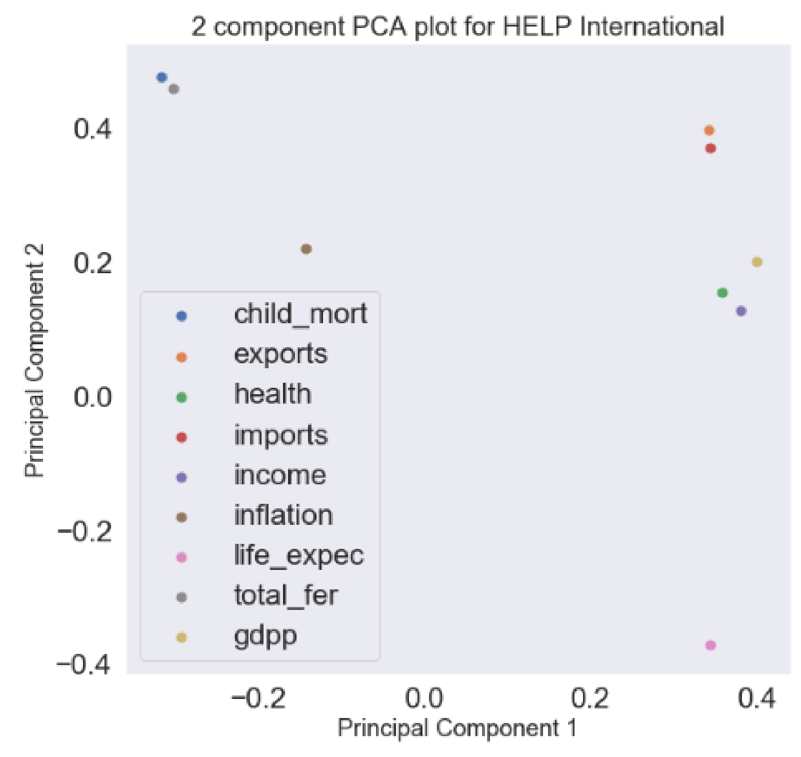
Inference:

Looks like only 4 components are enough to describe 94% of the variance in the dataset above.
We'll choose 4 components for our modeling.

Visualization for the Principal Components Selected from PCA

Since this is a 2-D plane, we can not plot the scatter plot between these four components directly.
Hence, we will go ahead and plot it for first 2 principal components which are most essential here



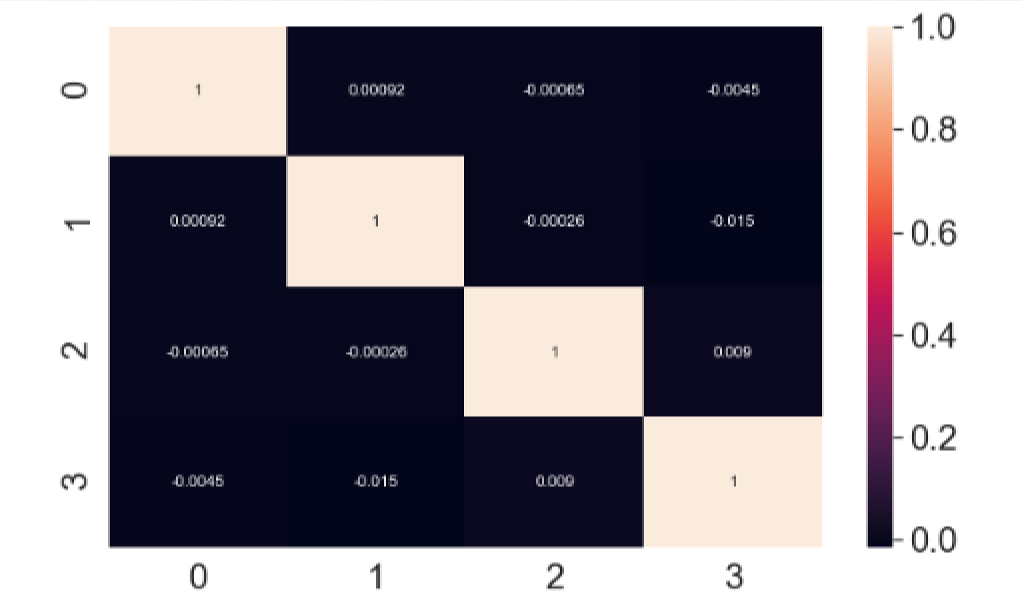


Analysis of the PC Plot between first two major components:

1. For component1 : The parameters like life expectancy (life_expec), income (income), health (health), gdpp, exports and imports are high for the "Principal Component1". Rest parameters are lesser.
2. For Principal Component 2 : The parameters like inflation, child mortality rate (child_mort) and (total_fer) total fertility rate are higher along with the above parameter, But life expectancy is lower.

fppt.com

Interpreting the result of the PCA

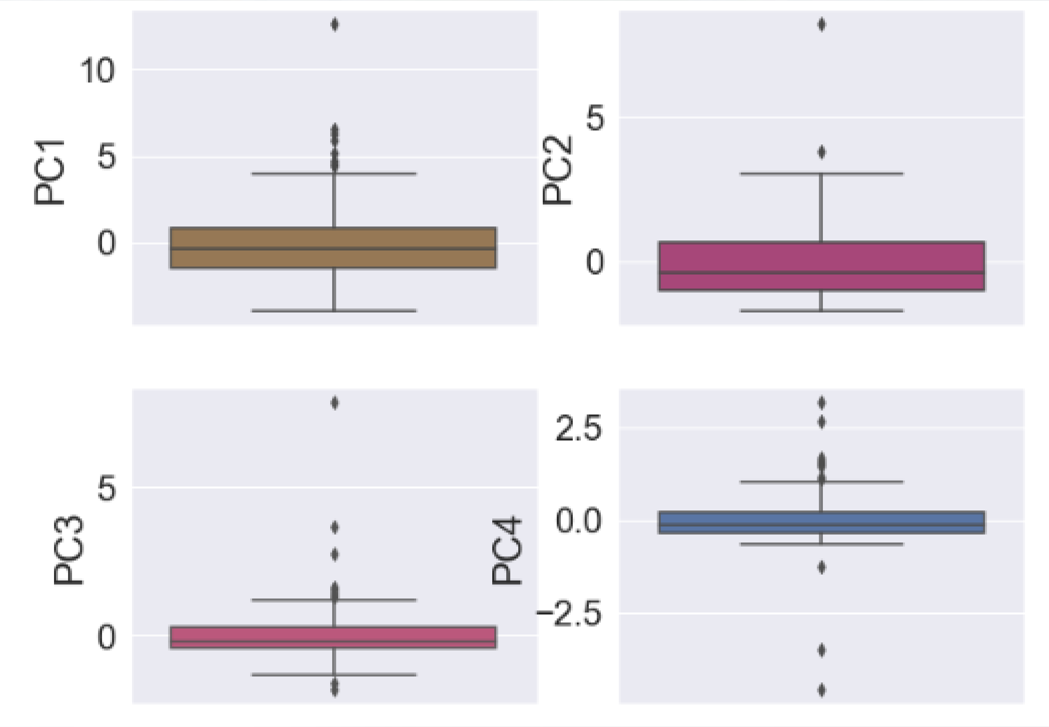


Inference:

As expected PCA has done quite well with very little to no correlation existing between the components as visible in the heatmap. This suggests that we can use these 4 components for next steps.

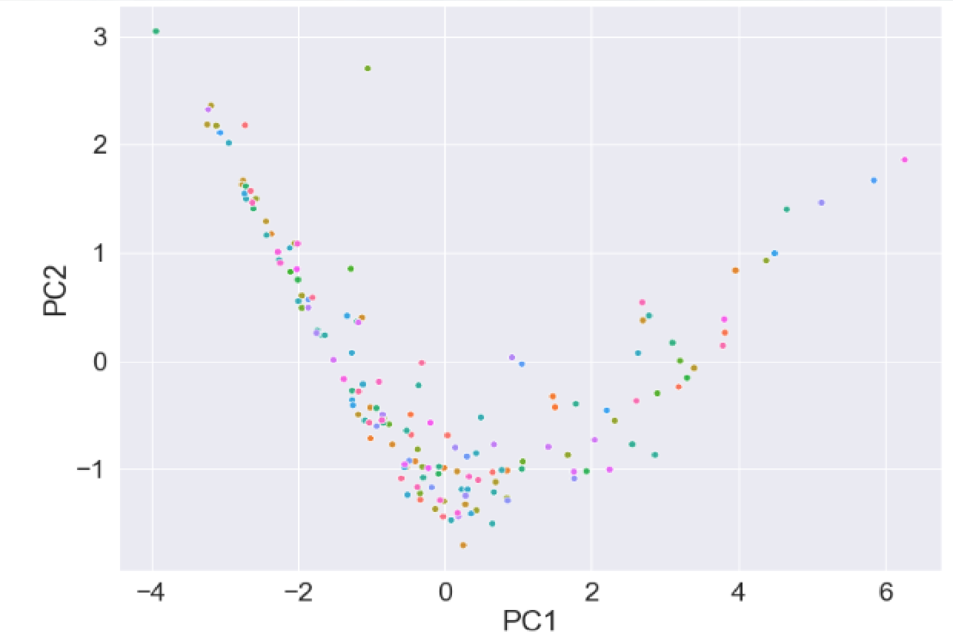
fppt.com

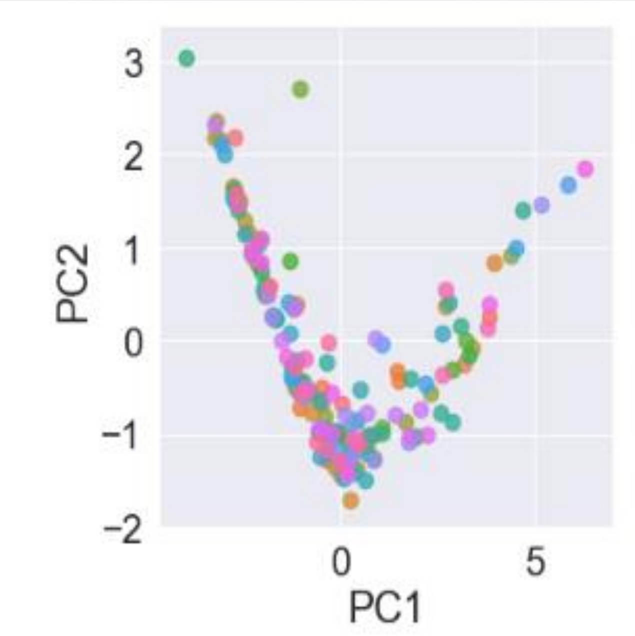
Performing Outlier Analysis after PCA



There are outliers in our data set for all the 4 components, which needs to be treated as clustering is sensitive to the presence of the outliers in the data.
So we can have eliminated the outliers in the data set.

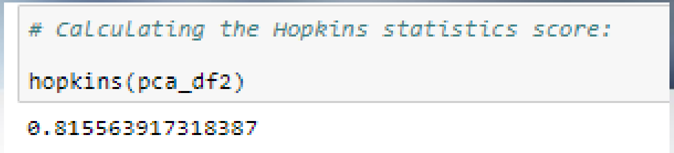
Visualizing the clusters created after PCA





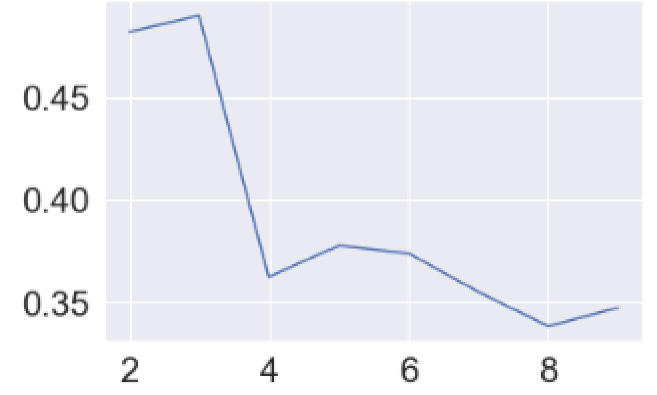
Some clusters are already visible to us, but they does not seems to be very clear now.
Let's go ahead and begin with the clustering process

Hopkins score, Silhouette Analysis and Elbow Curve for clustering: kMeans

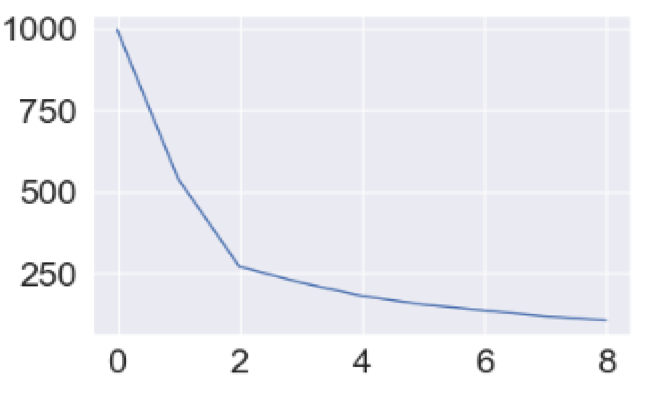


```
# Calculating the Hopkins statistics score:

hopkins(pca_df2)

0.815563917318387
```

Since the Hopkins score is ~0.81 , cluster tendency is very high for our data set.
This is as par the below Hopkins stats guidelines:
If the value is between {0.7, ..., 0.99}, it has a high tendency to cluster.

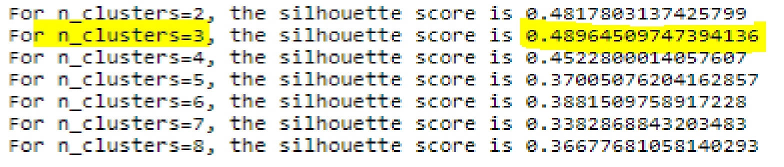


Silhouette Analysis



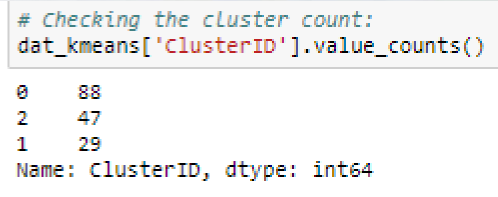
Elbow Curve

Inference:
From the above analysis we find that 3 seems to be a good number of clusters for K means algorithm.
From the elbow curve its suggesting 2 and 3 as the curve is bending at this point.
But the Silhouette analysis is suggesting as 3 , so we will go with the final model with **cluster number K=3 value**.



```
For n_clusters=2, the silhouette score is 0.4817803137425799
For n_clusters=3, the silhouette score is 0.4896450974739 4136
For n_clusters=4, the silhouette score is 0.4522800014057607
For n_clusters=5, the silhouette score is 0.37005076204162857
For n_clusters=6, the silhouette score is 0.3881509758917228
For n_clusters=7, the silhouette score is 0.3382868843203483
For n_clusters=8, the silhouette score is 0.36677681058140293
```

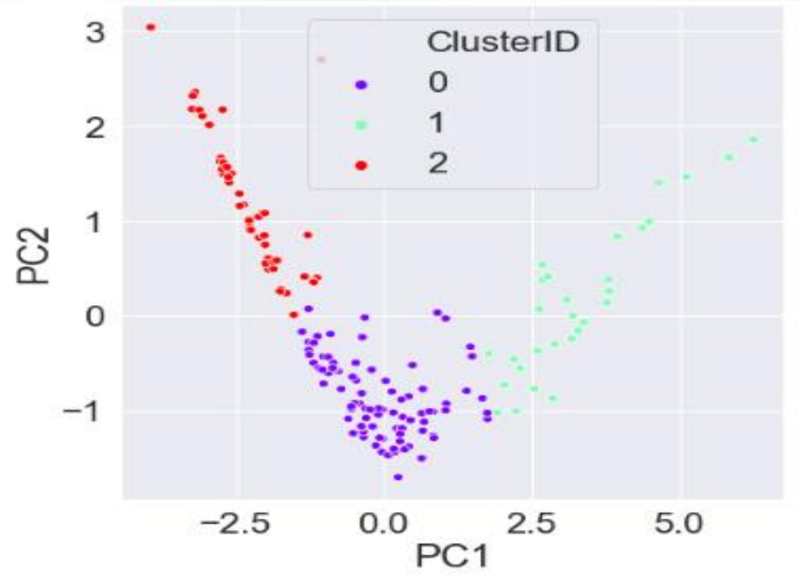
Understanding the cluster formed after PCA



```
# Checking the cluster count:
dat_kmeans['ClusterID'].value_counts()
```

```
0    88
2    47
1    29
Name: ClusterID, dtype: int64
```

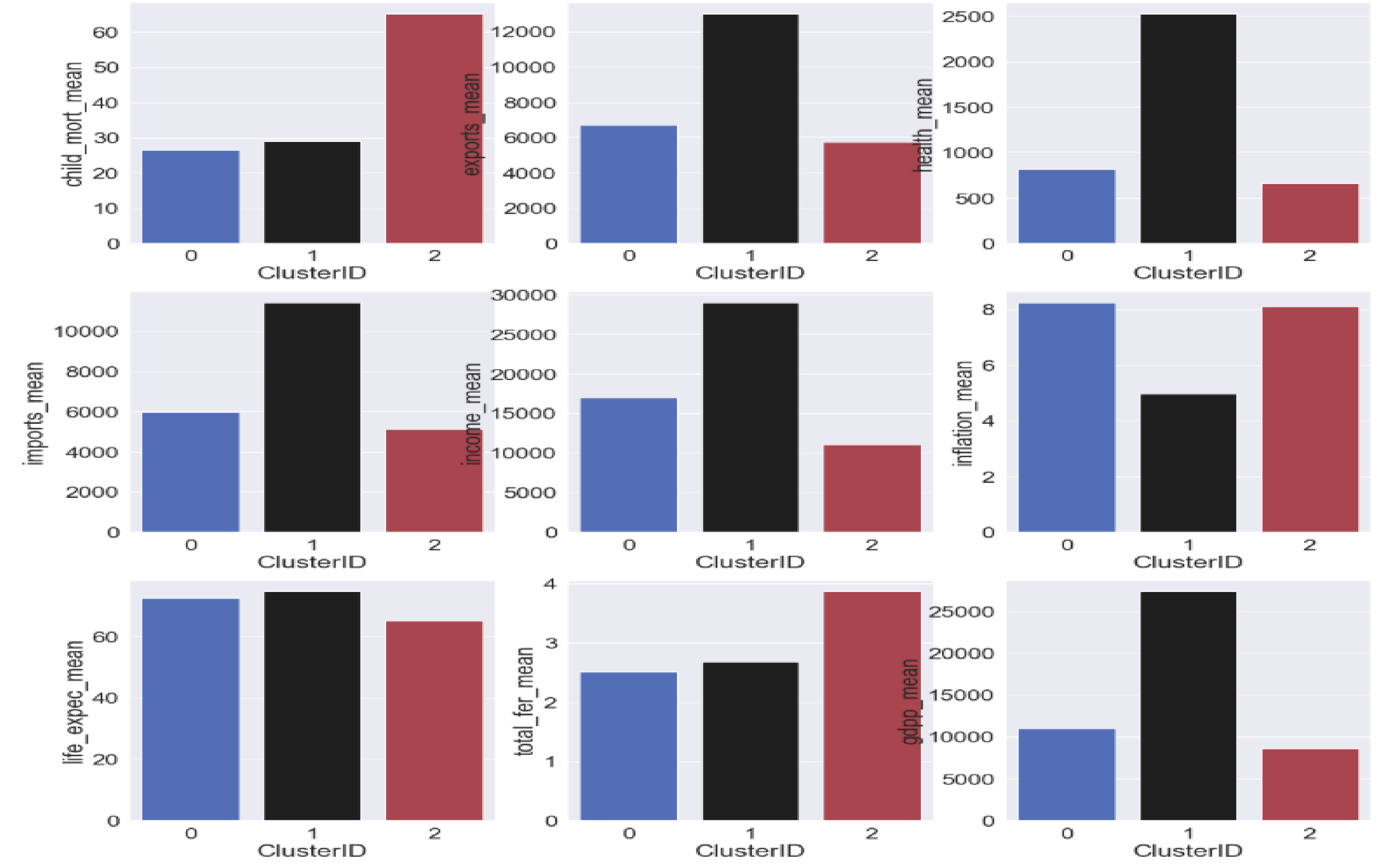
Hence there are 88 countries in cluster 1, 47 in cluster 2 and 29 in the 3rd cluster respectively.



So, basically we can see that there are three clusters which are formed here for all the countries.

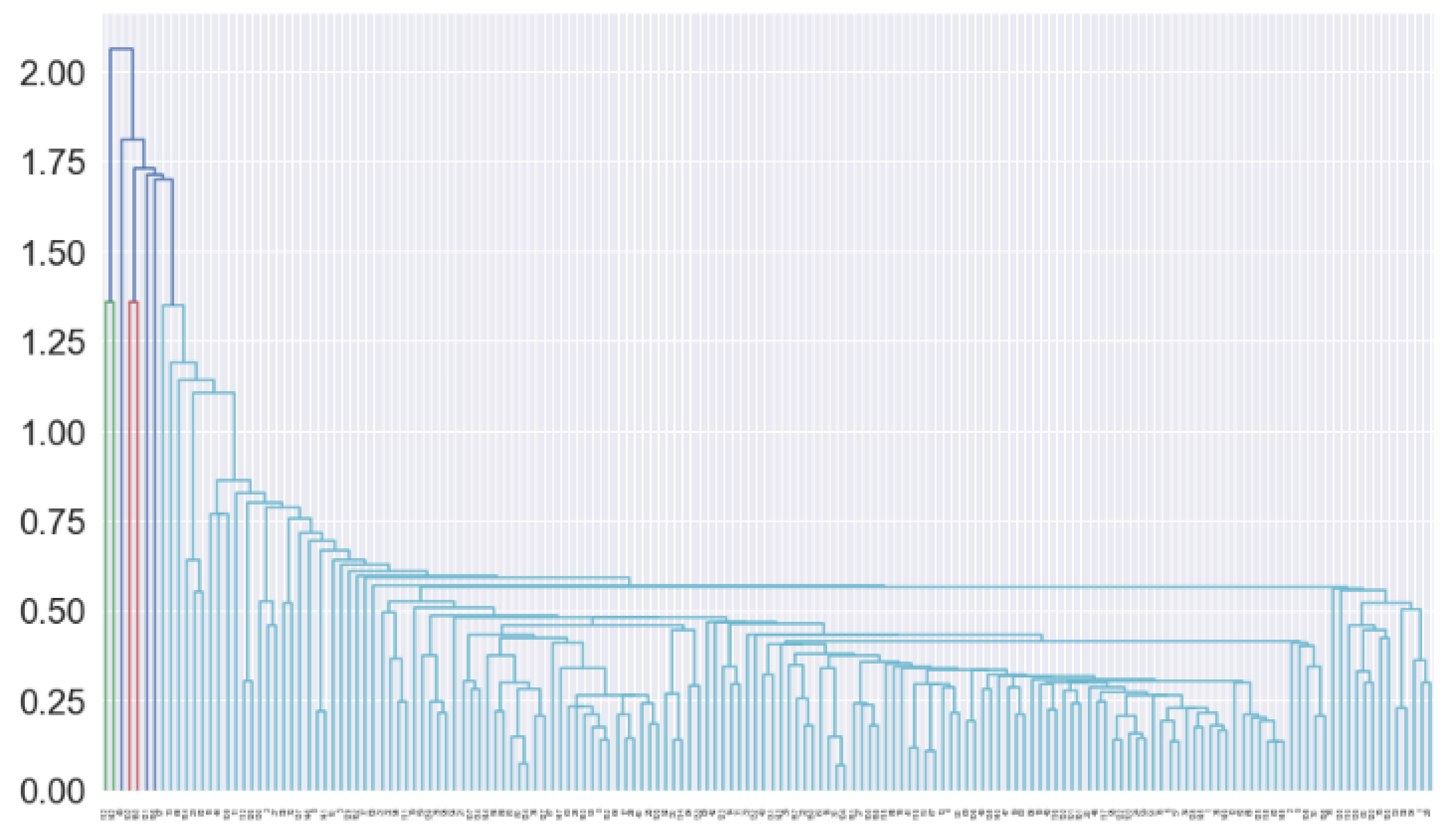
fppt.com

Performing Analysis for the cluster and other features formed : Kmeans Clustering



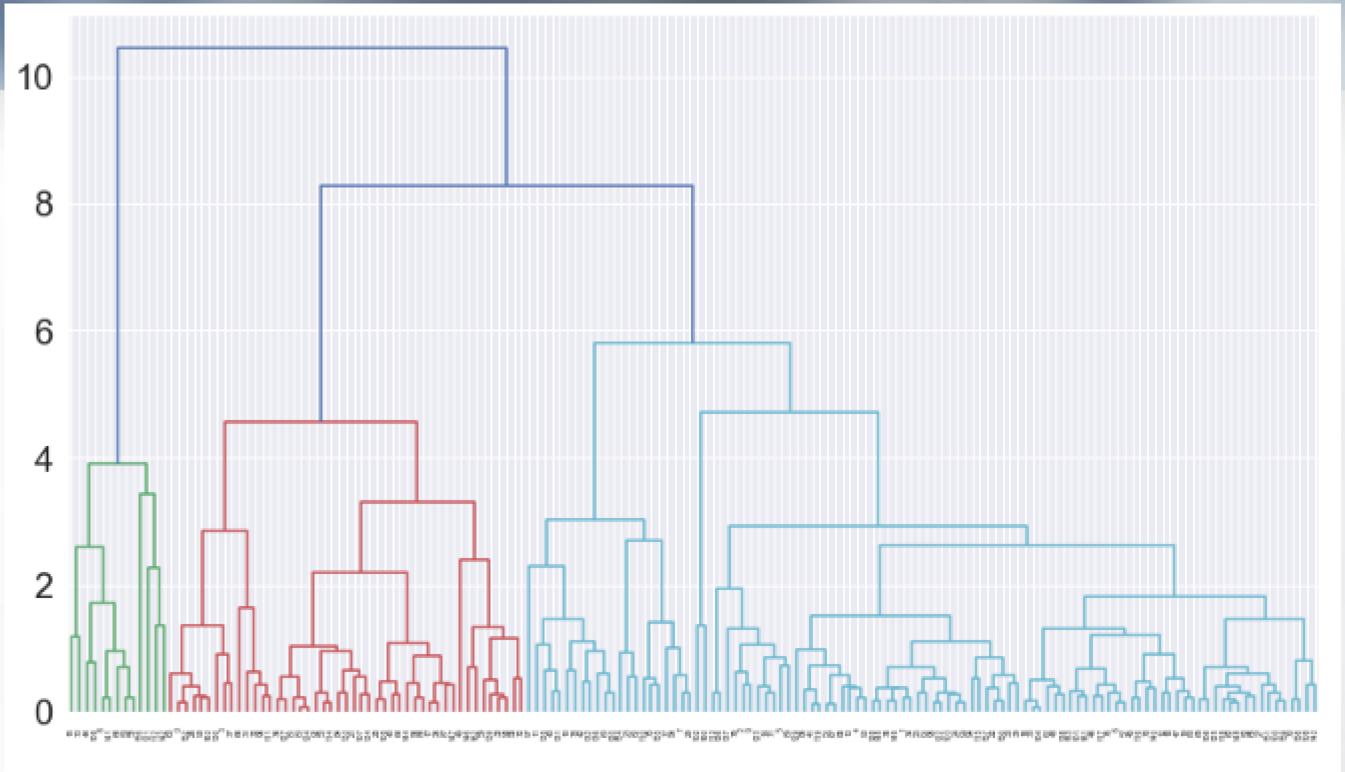
Here clusters are 0, 1 and 2 respectively.

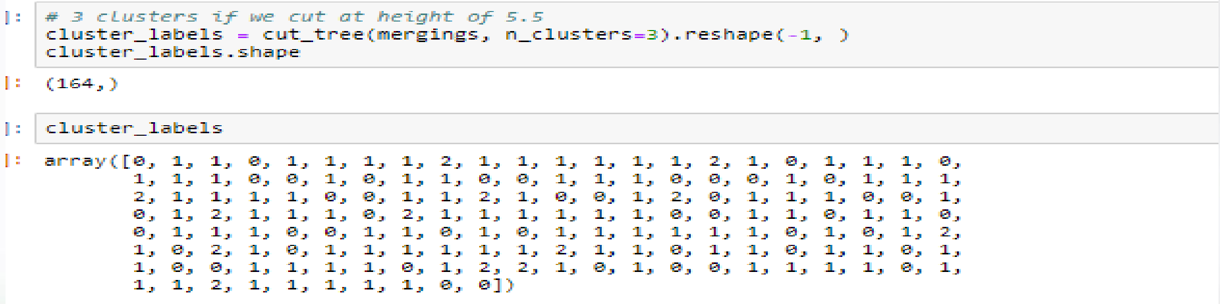
Hierarchical Clustering - Single linkage



Single linkage dendrogram

Hierarchical Clustering - Complete linkage

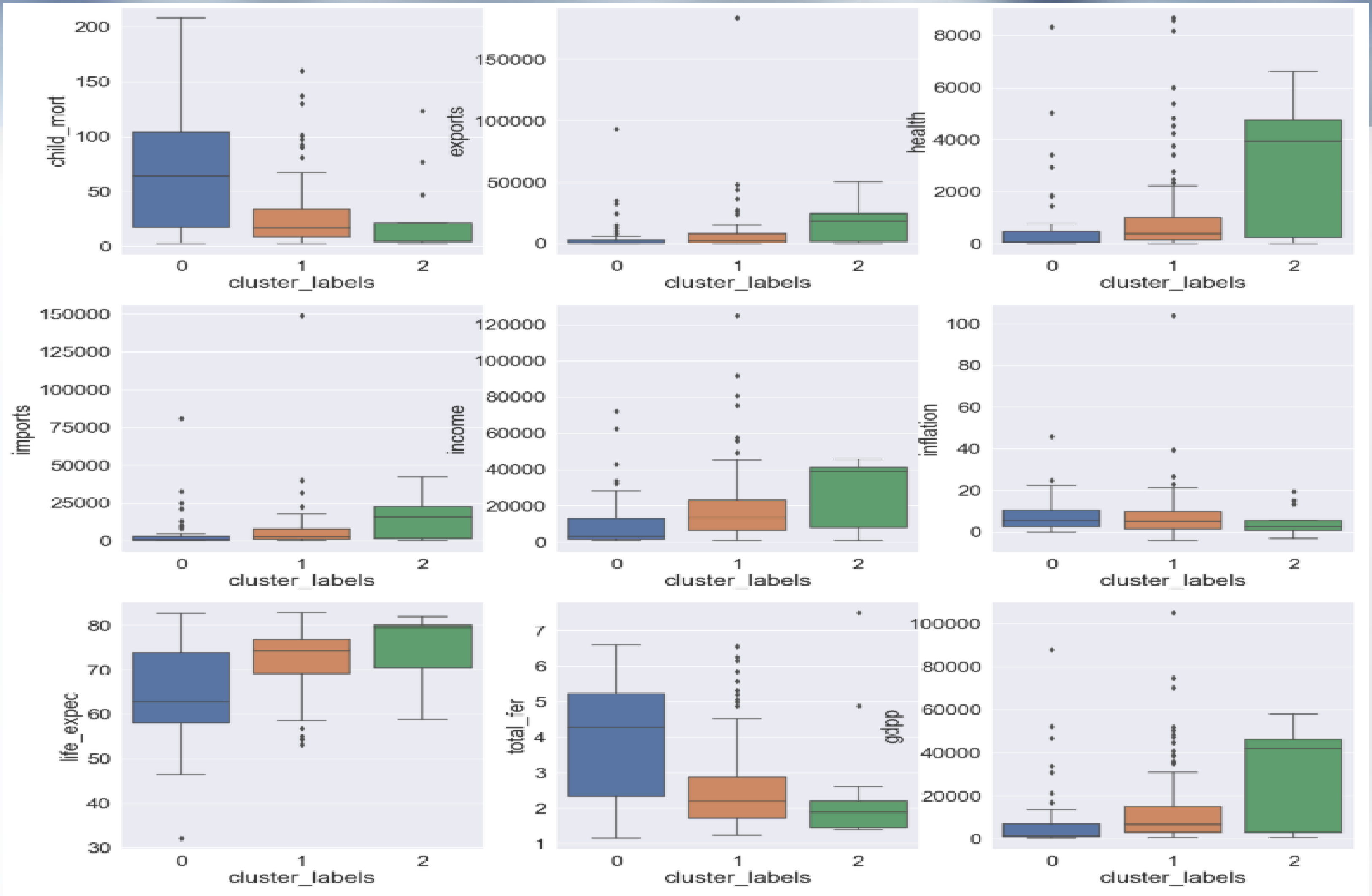




```
]: # 3 clusters if we cut at height of 5.5
   cluster_labels = cut_tree(mergings, n_clusters=3).reshape(-1, )
   cluster_labels.shape

]: (164,)

]: cluster_labels

]: array([0, 1, 1, 0, 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 2, 1, 0, 1, 1, 1, 0,
          1, 1, 1, 0, 0, 1, 0, 1, 1, 0, 0, 1, 1, 1, 0, 0, 0, 1, 0, 1, 1, 1,
          2, 1, 1, 1, 1, 0, 0, 1, 1, 2, 1, 0, 0, 1, 2, 0, 1, 1, 1, 0, 0, 1,
          0, 1, 2, 1, 1, 1, 0, 2, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 0, 1, 1, 0,
          0, 1, 1, 1, 0, 0, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 2,
          1, 0, 2, 1, 0, 1, 1, 1, 1, 1, 1, 2, 1, 1, 0, 1, 1, 0, 1, 1, 0, 1,
          1, 0, 0, 1, 1, 1, 1, 0, 1, 2, 2, 1, 0, 1, 0, 0, 1, 1, 1, 1, 0, 1,
          1, 1, 2, 1, 1, 1, 1, 1, 0, 0])
```

fppt.com

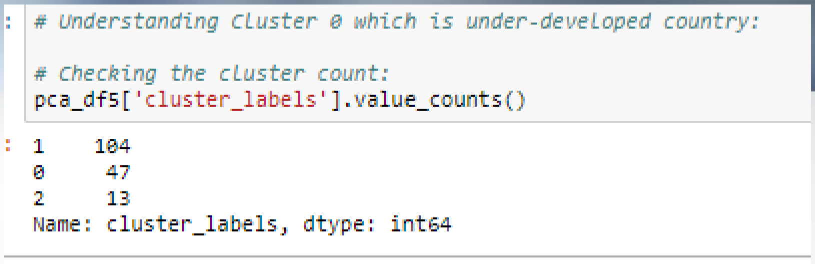
Visualization of Data after performing Hierarchical Clustering



Inference: As evident from the above box plot **cluster 0 represents the under-developed country** with least gdpp, life_expec, income, imports, health, exports etc. On the other hand features like child_mort, inflation, total_fer is highest for the countries under this cluster. Cluster 1 represents the developing country and Cluster 2 represents the developed country.

fppt.com

Understanding the clusters: Hierarchical Clustering

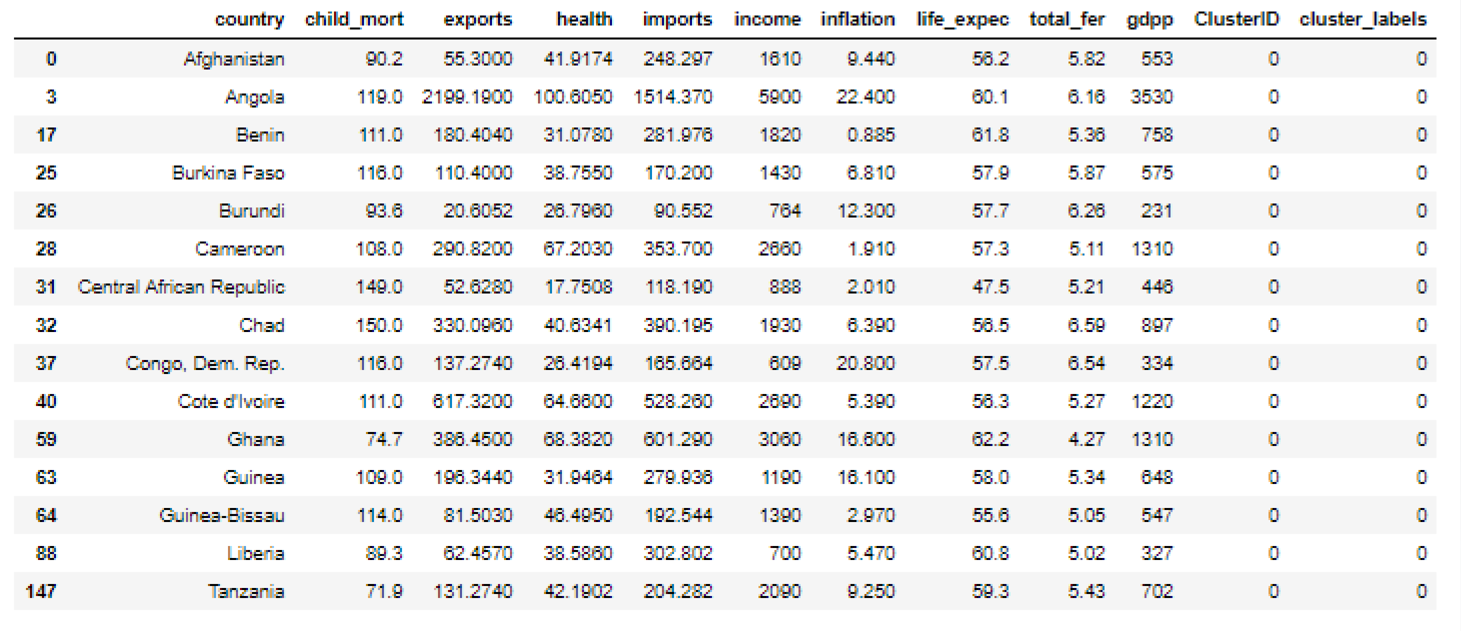


```
# Understanding Cluster 0 which is under-developed country:

# Checking the cluster count:
pca_df5['cluster_labels'].value_counts()
```

```
1    104
0     47
2     13
Name: cluster_labels, dtype: int64
```

So there are 47 countries which are undeveloped country and present in the cluster 0.

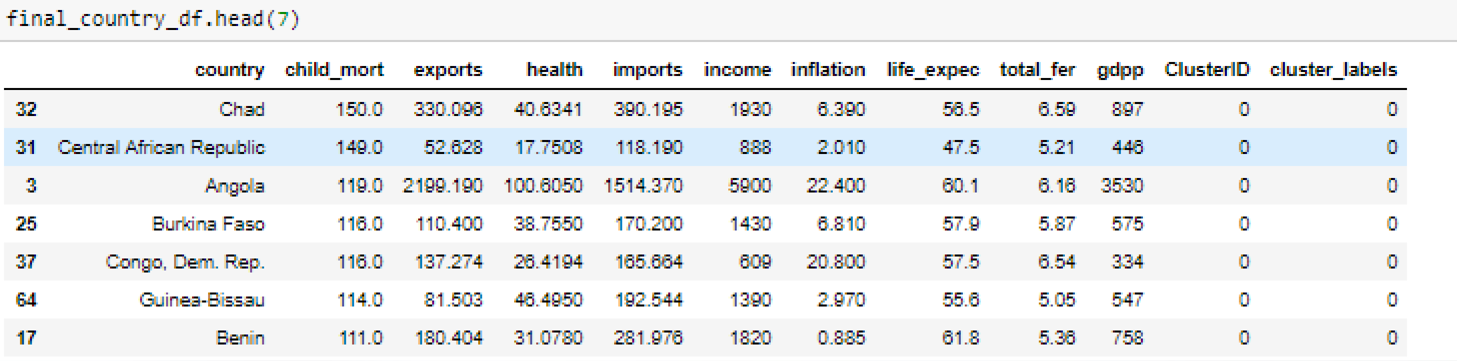
Checking the final data frame after binning to filter out the most under-developed countries in this under-developed cluster based on means of each feature:



	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	ClusterID	cluster_labels
0	Afghanistan	90.2	55.3000	41.9174	248.297	1610	9.440	56.2	5.82	553	0	0
3	Angola	119.0	2199.1900	100.6050	1514.370	5900	22.400	60.1	6.16	3530	0	0
17	Benin	111.0	180.4040	31.0780	281.976	1820	0.885	61.8	5.36	758	0	0
25	Burkina Faso	116.0	110.4000	38.7550	170.200	1430	6.810	57.9	5.87	575	0	0
26	Burundi	93.6	20.6052	26.7960	90.552	764	12.300	57.7	6.26	231	0	0
28	Cameroon	108.0	290.8200	67.2030	353.700	2660	1.910	57.3	5.11	1310	0	0
31	Central African Republic	149.0	52.6280	17.7508	118.190	888	2.010	47.5	5.21	446	0	0
32	Chad	150.0	330.0960	40.6341	390.195	1930	6.390	56.5	6.59	897	0	0
37	Congo, Dem. Rep.	116.0	137.2740	26.4194	165.664	609	20.800	57.5	6.54	334	0	0
40	Cote d'Ivoire	111.0	617.3200	64.6600	528.260	2690	5.390	56.3	5.27	1220	0	0
59	Ghana	74.7	386.4500	68.3820	601.290	3060	16.600	62.2	4.27	1310	0	0
63	Guinea	109.0	196.3440	31.9464	279.936	1190	16.100	58.0	5.34	648	0	0
64	Guinea-Bissau	114.0	81.5030	46.4950	192.544	1390	2.970	55.6	5.05	547	0	0
88	Liberia	89.3	62.4570	38.5860	302.802	700	5.470	60.8	5.02	327	0	0
147	Tanzania	71.9	131.2740	42.1902	204.282	2090	9.250	59.3	5.43	702	0	0

fppt.com

Summary and Recommendation



```
final_country_df.head(7)
```

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	ClusterID	cluster_labels
32	Chad	150.0	330.096	40.6341	390.195	1930	6.390	56.5	6.59	897	0	0
31	Central African Republic	149.0	52.628	17.7508	118.190	888	2.010	47.5	5.21	446	0	0
3	Angola	119.0	2199.190	100.6050	1514.370	5900	22.400	60.1	6.16	3530	0	0
25	Burkina Faso	116.0	110.400	38.7550	170.200	1430	6.810	57.9	5.87	575	0	0
37	Congo, Dem. Rep.	116.0	137.274	26.4194	165.664	609	20.800	57.5	6.54	334	0	0
64	Guinea-Bissau	114.0	81.503	46.4950	192.544	1390	2.970	55.6	5.05	547	0	0
17	Benin	111.0	180.404	31.0780	281.976	1820	0.885	61.8	5.36	758	0	0

As we can see from the above results all these underdeveloped countries falls in the same clusters in both kMeans and Hierarchial clustering process.

Summary: Below are the countries which are selected for the recommendation to HELP International and they are in direst need of financial aid:

Chad
Central African Republic
Angola
Burkina Faso
Congo, Dem. Rep.
Benin

fppt.com