Clustering and PCA Assignment

Arihant Sethia

Problem Statement

- To cluster the countries by factors given in the dataset (socio-economic and health factors) that determine the overall development of the country.
- To do Dimensionality Reduction using PCA.
- Identify the cluster of countries which are in most need of funding.

Assumptions:

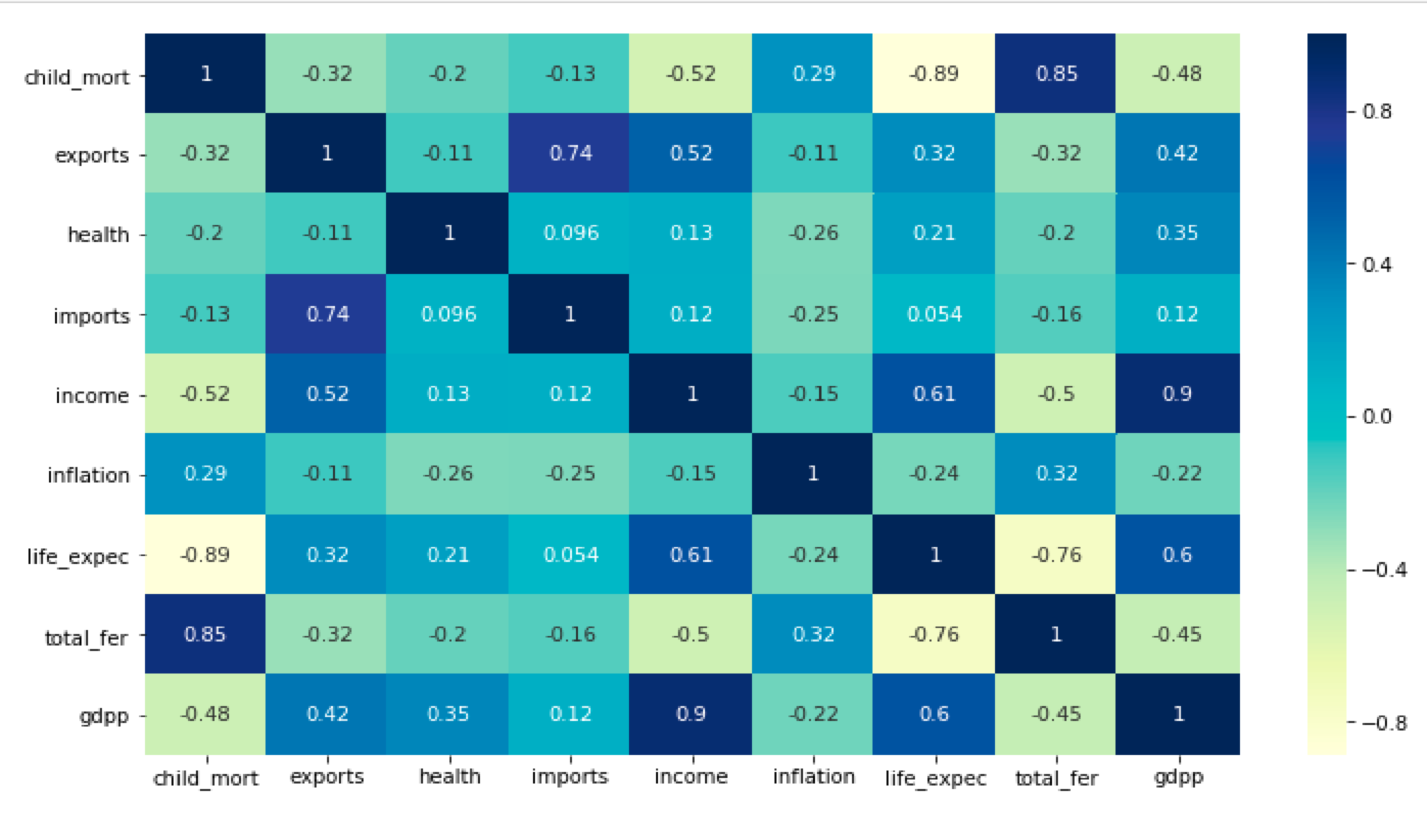
- All the data given in the sheet is correct.
- All countries need to be analyzed carefully, even if they categorize as outliers.
- Outlier data points can be added to clusters by identifying the nearest cluster centers.
- We only need to identify the cluster of countries which is in most need of funding.

Problem Solving Methodology

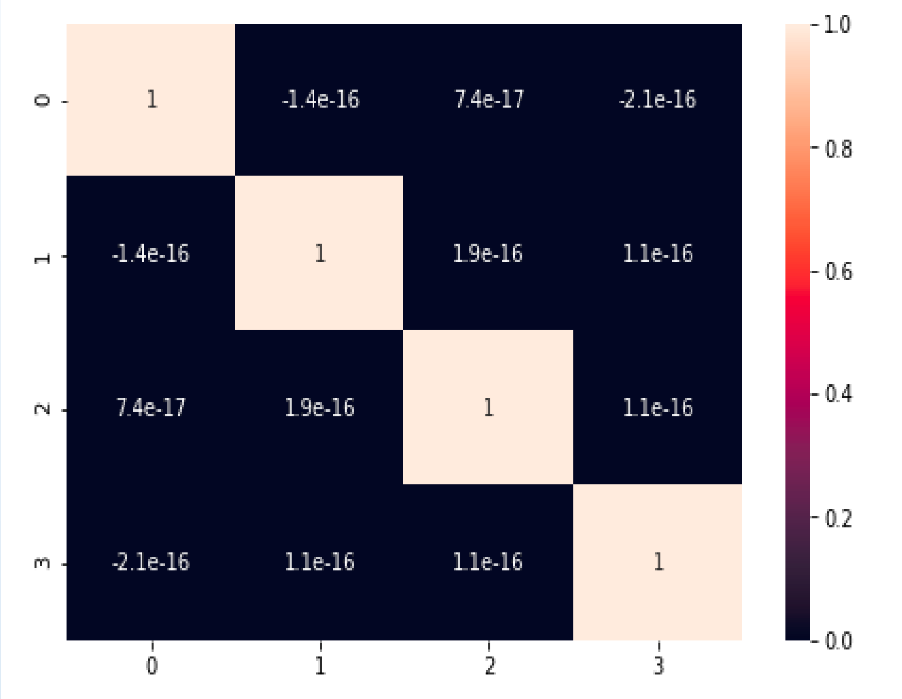
1. Data Exploration.
2. Identify co-relation among different columns.
3. Scale the data and do Principal Component Analysis.
4. Based on screeplot, identify the number of Principal Components which explains > 85% of variance.
5. Plot the co-relation map again to check if co-linearity is handled by doing above step.
6. Remove the outliers from the dataset using IQR method.
7. Compute Hopkins Score, to check for clustering.
8. Compute Silhoutte score and plot elbow curve to find optimum number of K for K-Means.
9. Assign cluster_ids to the dataset using K-Means.
10. Assign cluster_id to the outlier data based on nearest cluster center.
11. Joining the dataset having cluster_ids with original dataset (having original features).
12. Drop the PCs. Analyse the clusters based on original features.
13. Create hierarchical clustering on the PCA data and add original features back to them.
14. Based on above clustering and analysing the features identify the cluster of countries which is in most need of funding.

Need for doing PCA

- We can see there are 9 variables and some of them are highly co-related.
- By doing PCA we will reduce the dimension and also the final features will have almost zero correlation.



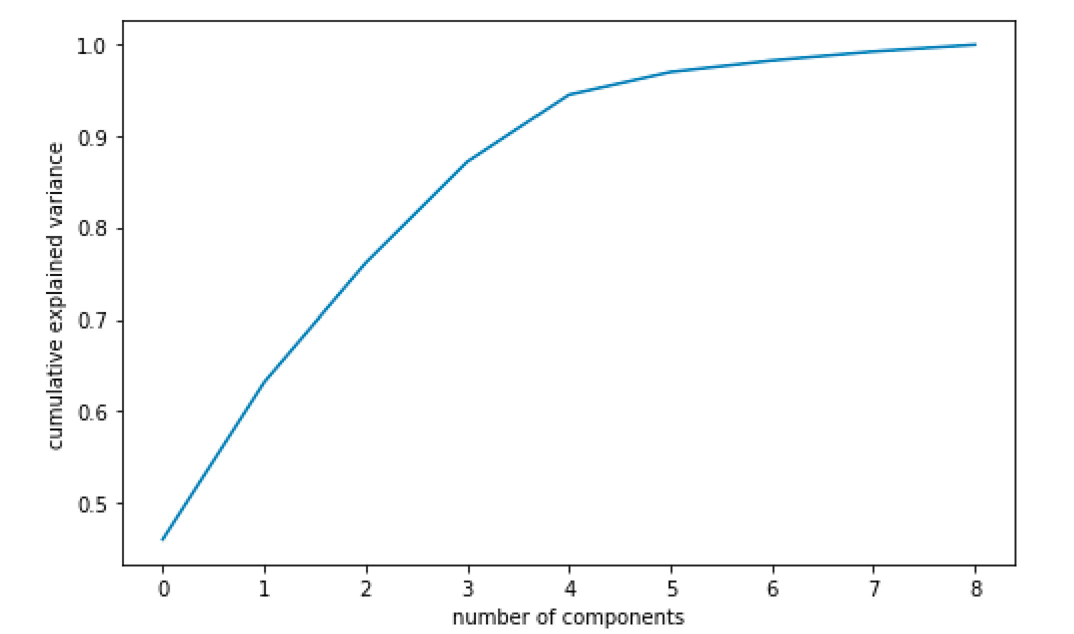
	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
child_mort	1	-0.32	-0.2	-0.13	-0.52	0.29	-0.89	0.85	-0.48
exports	-0.32	1	-0.11	0.74	0.52	-0.11	0.32	-0.32	0.42
health	-0.2	-0.11	1	0.096	0.13	-0.26	0.21	-0.2	0.35
imports	-0.13	0.74	0.096	1	0.12	-0.25	0.054	-0.16	0.12
income	-0.52	0.52	0.13	0.12	1	-0.15	0.61	-0.5	0.9
inflation	0.29	-0.11	-0.26	-0.25	-0.15	1	-0.24	0.32	-0.22
life_expec	-0.89	0.32	0.21	0.054	0.61	-0.24	1	-0.76	0.6
total_fer	0.85	-0.32	-0.2	-0.16	-0.5	0.32	-0.76	1	-0.45
gdpp	-0.48	0.42	0.35	0.12	0.9	-0.22	0.6	-0.45	1



	0	1	2	3
0	1	-1.4e-16	7.4e-17	-2.1e-16
1	-1.4e-16	1	1.9e-16	1.1e-16
2	7.4e-17	1.9e-16	1	1.1e-16
3	-2.1e-16	1.1e-16	1.1e-16	1

After doing the PCA and taking 4 components, we can see that there is almost zero co-relation between all the PCs.

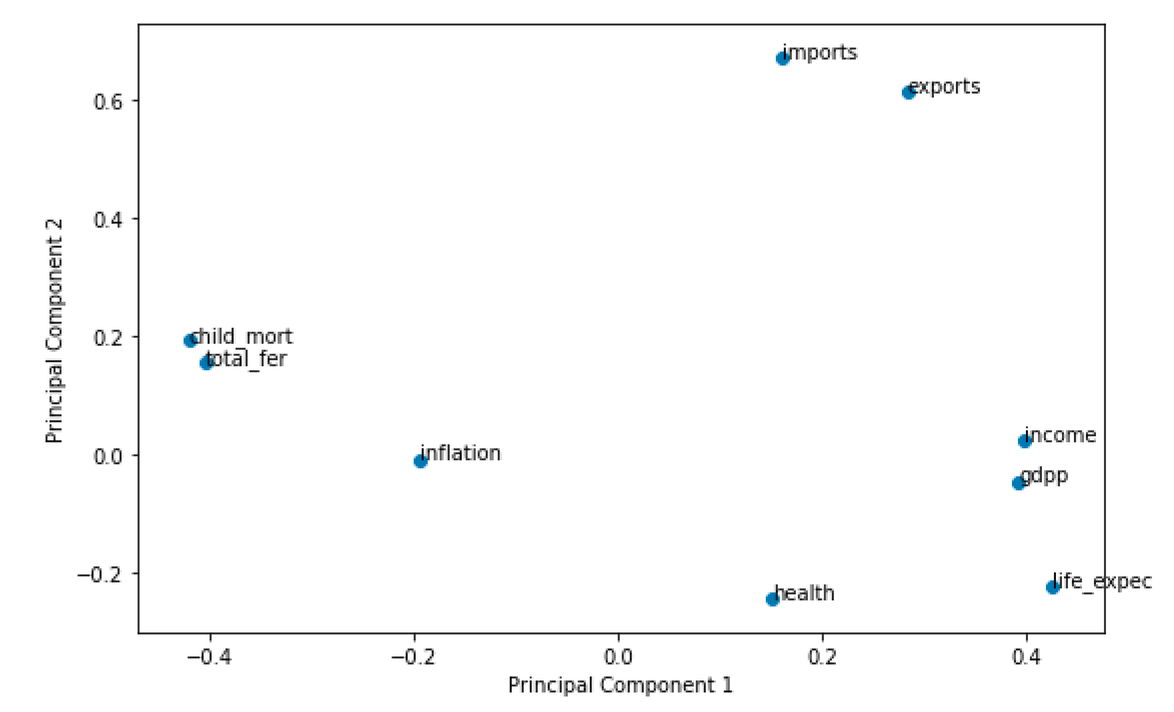
PCA



- From Above Scree Plot we can see that 4 principal components explains around 87% of the variance in the dataset. So, we will take 4 PCs.

PC1 vs PC2 plot vs original features:

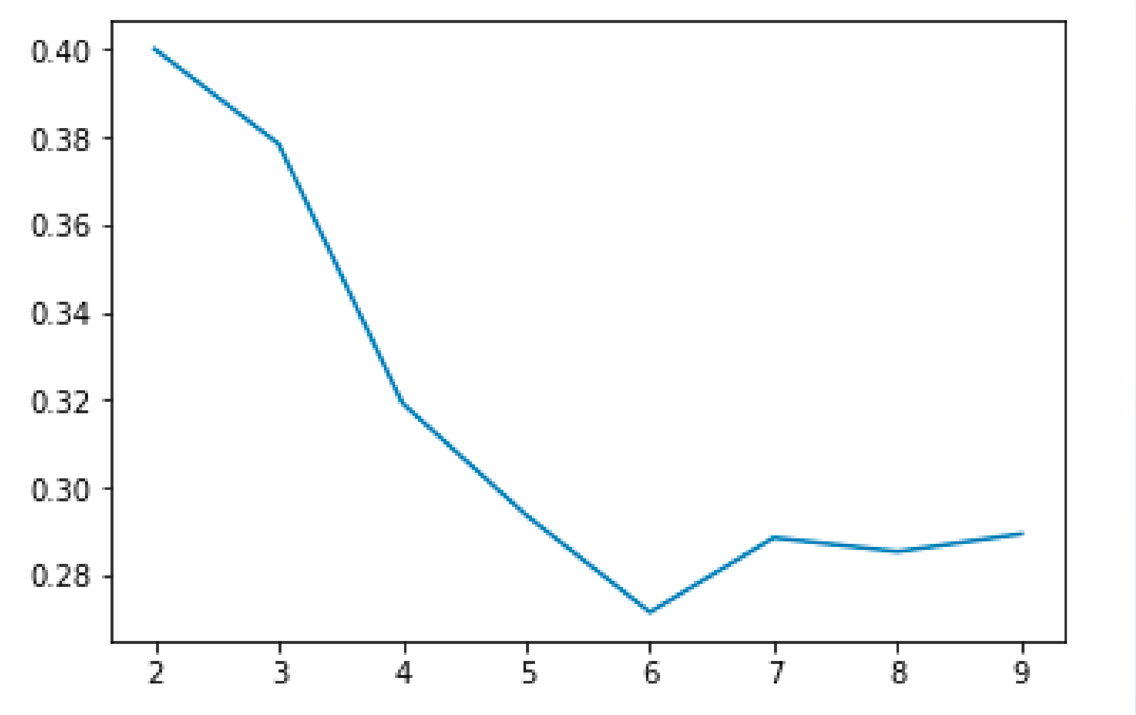
- We can see that export and import both have high PC1 and PC2 values.
- Child mortality and total fertility both have high PC1(negative) and PC2.
- Income and GDPP have high PC1 value.

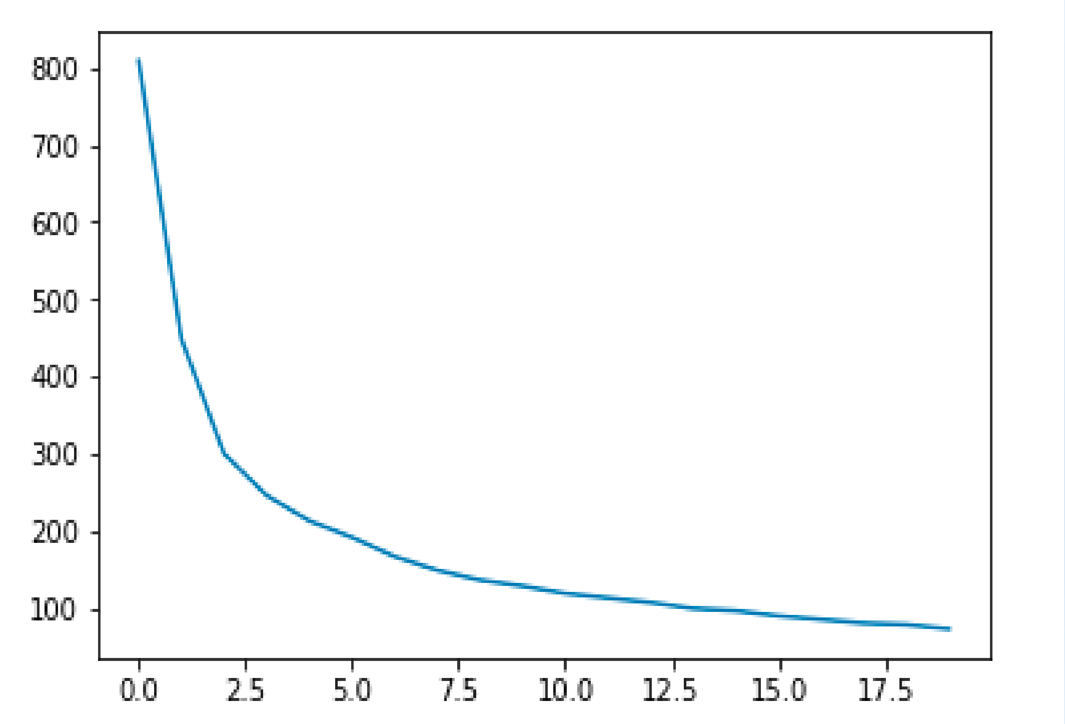


K-Means

Elbow Curve:

- We can take cluster size = 5.
- By further increasing cluster size beyond 5 there isn’t significant change in Sum of Squared Errors on y axis.



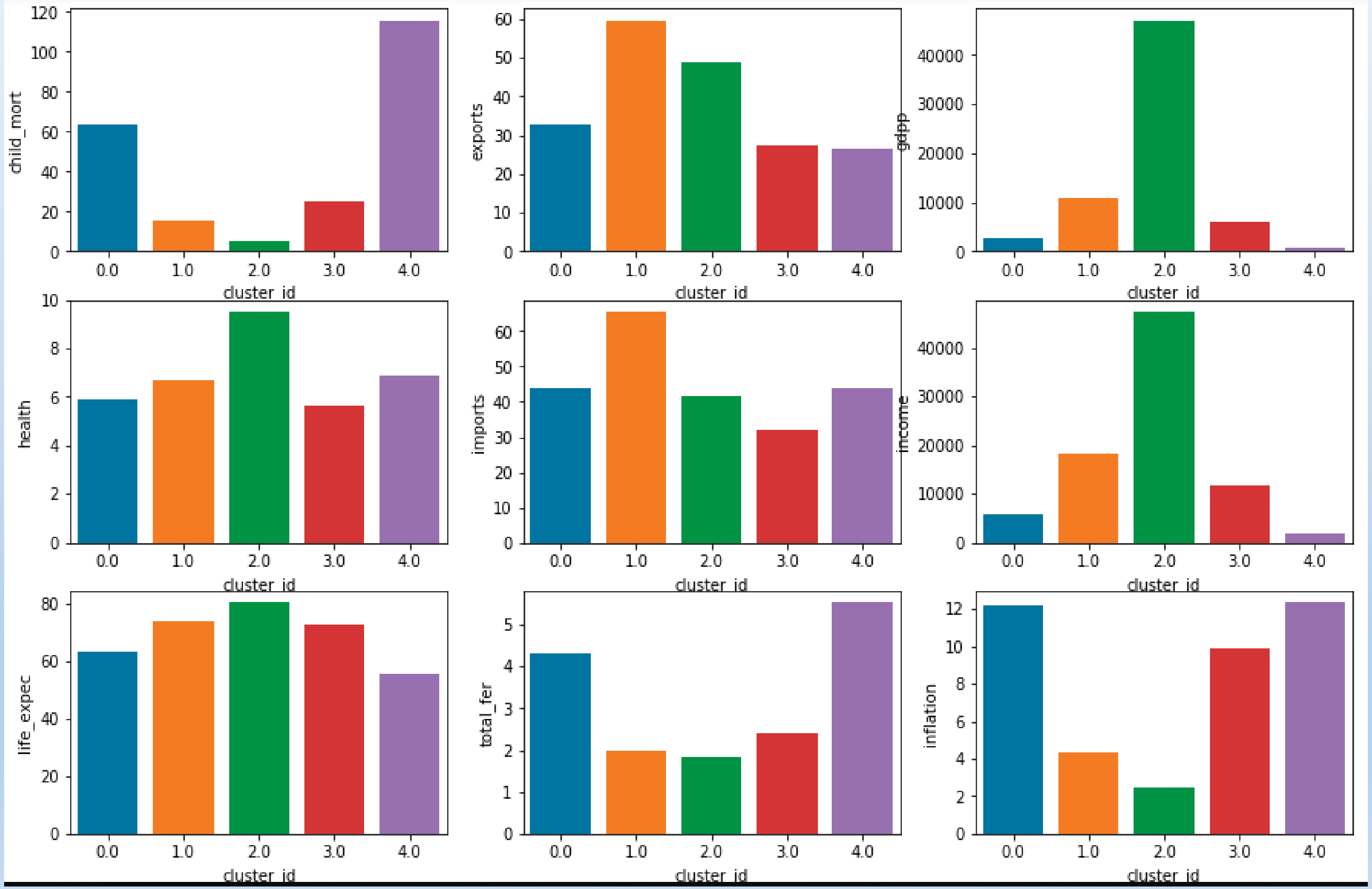


Silhoutte Coefficient:

This plot gives us a cluster size to choose. We can choose cluster size = 5. Although it is not the highest value but based on elbow curve, analyzing the data and silhouette score we choose cluster size = 5.

K-Means Clusters

- All the original features of the dataset plotted against the means of that feature per-cluster.
- We can observe that each cluster is significantly different from others based on all these features.
- We have also plotted scatter plots for individual values in the clusters and can differentiation between clusters based on them.



Observations/ Conclusions

Based on the plots and final dataset we can draw following conclusions:

1. Countries falling in the cluster_id = 4 have highest child mortality rates, lowest GDPP, lowest income, high inflation, lowest life expectancy, highest fertility rate and lowest exports among all clusters. This cluster contains the countries which is in most need of funding.

2. Cluster_id = 4 is followed by cluster_id = 0. These countries also have high child mortality rates, low GDPP, low income, high inflation, low life expectancy, high fertility rate and low exports. This cluster can also use funding if it is available after investing in cluster_id = 4.

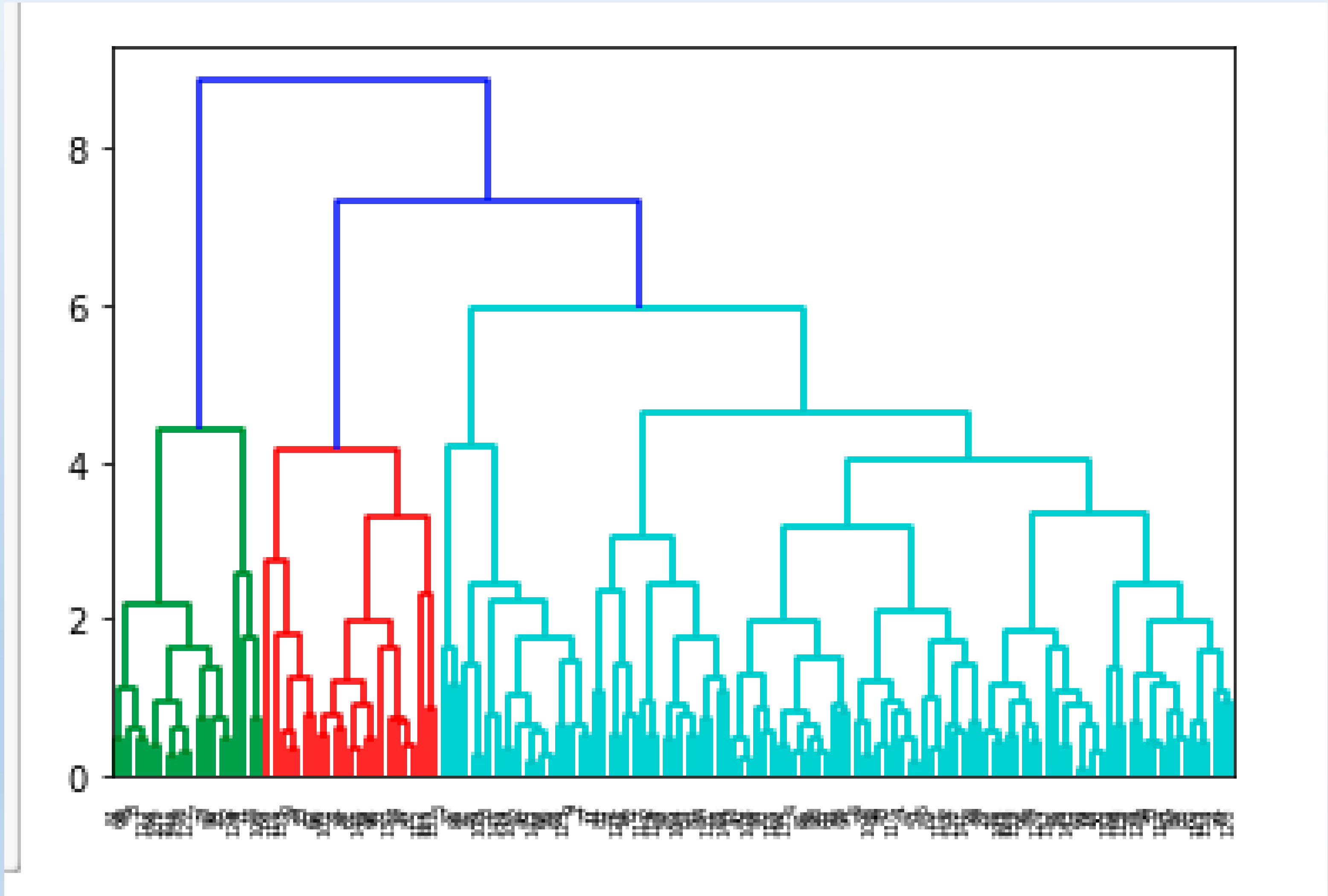
3. Cluster_id = 0 is followed by cluster_id = 3. These countries have lower child mortality rate, inflation and total fertility rate than cluster_id = 0 and 4 but still higher than cluster_id = 1 and 2. They have higher GDPP, income and life expectancy than cluster_id = 0 and 4 but still higher than cluster_id = 1 and 2.

4. Countries in cluster_id = 1 have low child mortality rates, high exports and imports, high GDP and income, low inflation and total fertility and high life expectancy. However, there is significant difference in all these numbers with other clusters.

5. Countries in cluster_id = 2 have lowest child_mortality rates, lowest inflation, low fertility. They have high export, import, highest GDPP, highest income and life expectancy among all clusters. This cluster contains developed countries.

Hierarchical Clustering

- If we cut the dendrogram somewhere between distance 4 and 6 we would get 5 clusters.
- We get similar result after creating cluster from hierarchical clustering and K-Means clustering.



Clusters Observation (from K-Means)

Countries	Conclusion
Botswana, Comoros, Congo, Rep., Equatorial Guinea, Eritrea, Gabon, Gambia, Ghana, Iraq, Kenya, Kiribati, Lao, Madagascar, Mauritania, Mongolia, Namibia, Pakistan, Rwanda, Senegal, Solomon Islands, South Africa, Sudan, Tajikistan, Tanzania, Timor-Leste, Vanuatu, Yemen	Cluster_ID = 0 This cluster of countries can also use funding if any funding is left after investing in cluster_ID = 4.
Antigua and Barbuda, Bahamas, Bahrain, Barbados, Belarus, Belize, Bhutan, Bosnia and Herzegovina, Bulgaria, Cambodia, Cape Verde, Croatia, Cyprus, Czech Republic, Estonia, Fiji, Georgia, Guyana, Hungary, Jordan, Kyrgyz Republic, Latvia, Lebanon, Libya, Lithuania, Macedonia, FYR, Malaysia, Maldives, Malta, Mauritius, Micronesia, Fed. Sts., Moldova, Montenegro, Oman, Panama, Paraguay, Poland, Serbia, Seychelles, Singapore, Slovak Republic, South Korea, Thailand, Tunisia, Turkmenistan, Ukraine, United Arab Emirates, Vietnam	Cluster_ID = 1 Developing Countries This cluster has higher GDPP, Income, export and import than cluster_id = 3.
Australia, Austria, Belgium, Brunei, Canada, Denmark, Finland, France, Germany, Greece, Iceland, Ireland, Israel, Italy, Japan, Kuwait, Luxembourg, Netherlands, New Zealand, Norway, Portugal, Qatar, Slovenia, Spain, Sweden, Switzerland, United Kingdom, United States	Cluster_ID = 2 Most developed countries. These don’t need any funding.
Albania, Algeria, Argentina, Armenia, Azerbaijan, Bangladesh, Bolivia, Brazil, Chile, China, Colombia, Costa Rica, Dominican Republic, Ecuador, Egypt, El Salvador, Grenada, Guatemala, India, Indonesia, Iran, Jamaica, Kazakhstan, Morocco, Myanmar, Nepal, Peru, Philippines, Romania, Russia, Samoa, Saudi Arabia, Sri Lanka, St. Vincent and the Grenadines, Suriname, Tonga, Turkey, Uruguay, Uzbekistan, Venezuela	Cluster_ID = 3 Developing Countries. This cluster has lower GDPP, Income, export and import than cluster_id = 1.
Afghanistan, Angola, Benin, Burkina Faso, Burundi, Cameroon, Central African Republic, Chad, Congo, Dem. Rep., Cote d'Ivoire, Guinea, Guinea-Bissau, Haiti, Lesotho, Liberia, Malawi, Mali, Mozambique, Niger, Nigeria, Sierra Leone, Togo, Uganda, Zambia	Cluster_ID = 4 This cluster of countries is in most need of funding. We need to invest in these to improve living conditions.