

Clustering and PCA Assignment

By :
Naresh Arora

Agenda



Discuss about objective of Assignment for given problem statement



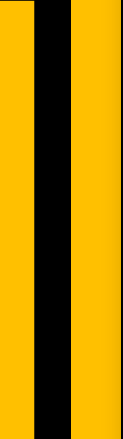
Data Processing



Clustering Process



Conclusion



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 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.

Objectives



Provided summary level insight into Country data using Jupyter notebook.



Uncovered underlying patterns and used clustering algorithms to provide recommendations.



To categorize the countries using socio-economic and health factors and suggest the list of same which CEO needs to focus most.

Tools used for analysis



JUPYTER NOTEBOOK

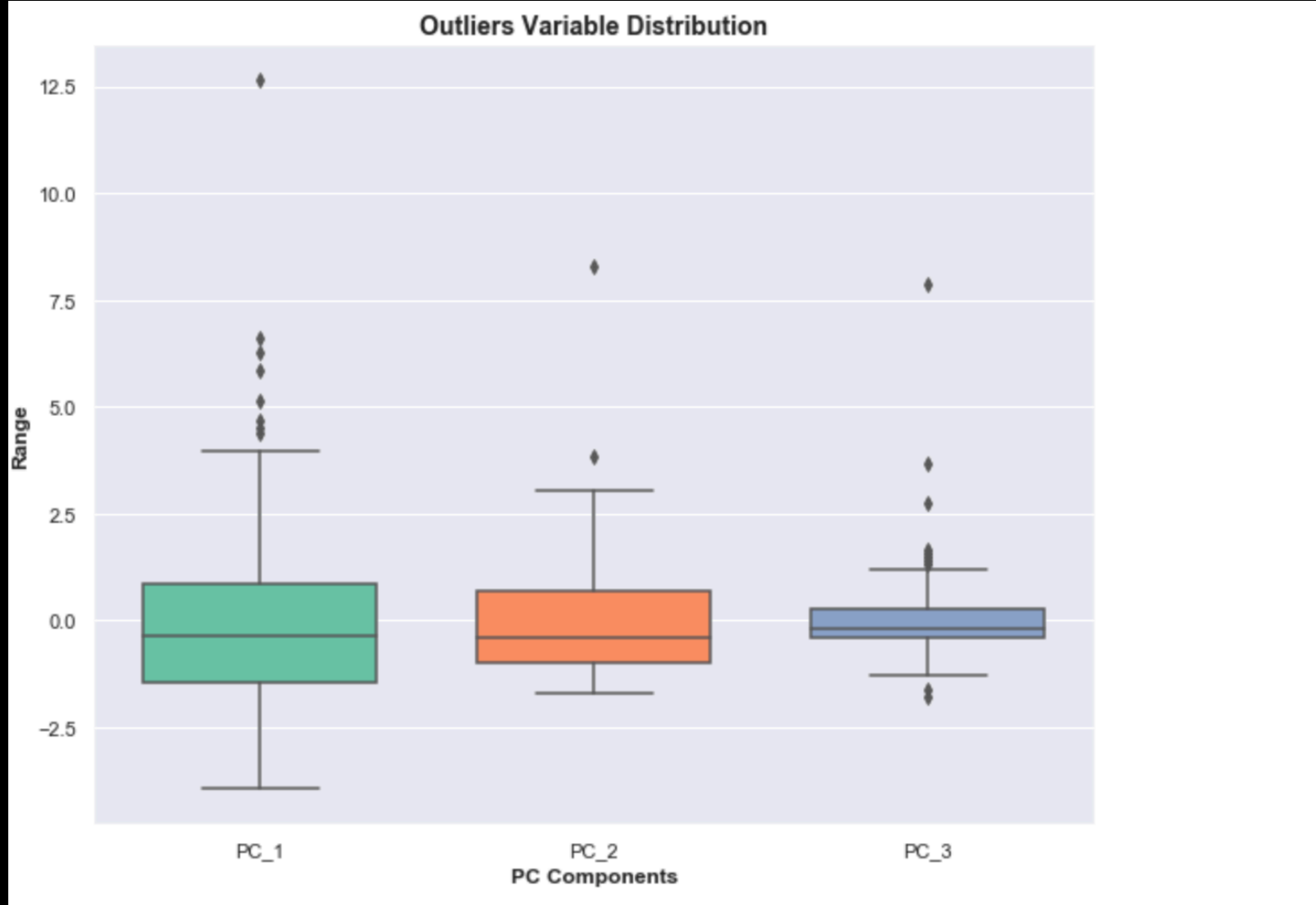


MICROSOFT EXCEL

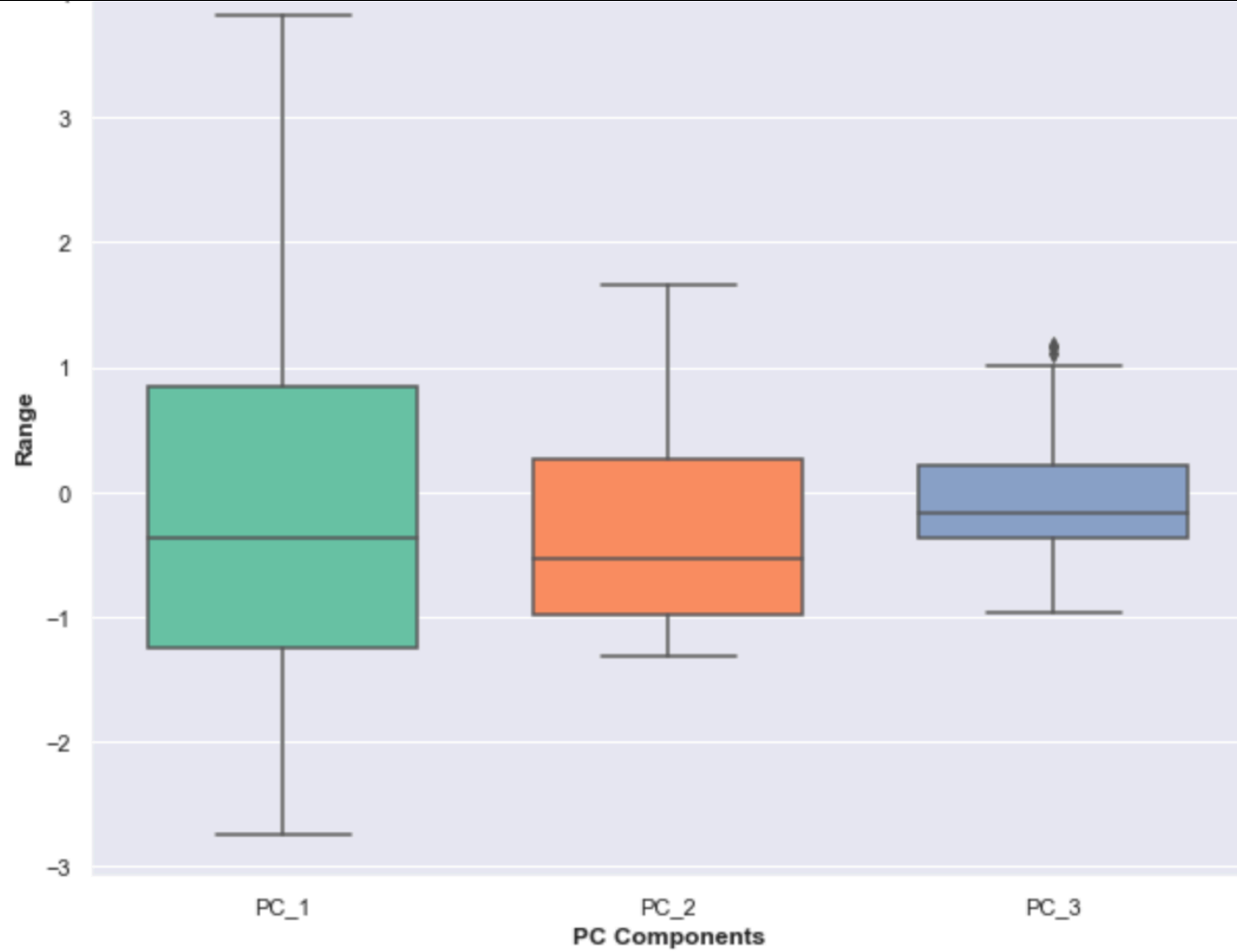


Discuss Data Model

- There are no null values in the dataset provided.
- There are no duplicate values for the country.
- Outliers data is analyzed and box plots are created to show data with and without outliers.
- Standardization of data using PCA(Principal Component Analysis)



With
Outliers

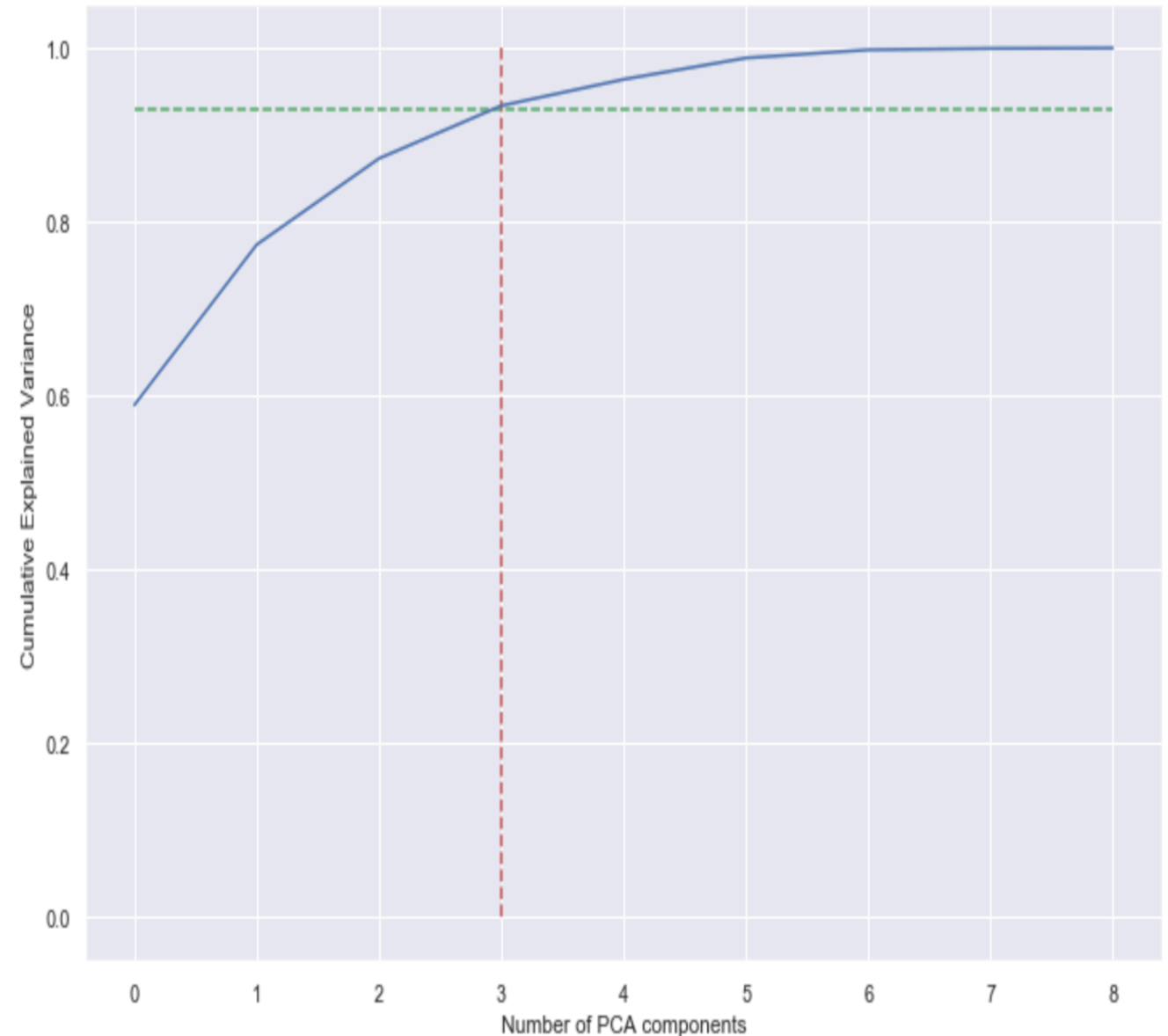


Without
Outliers

PCA

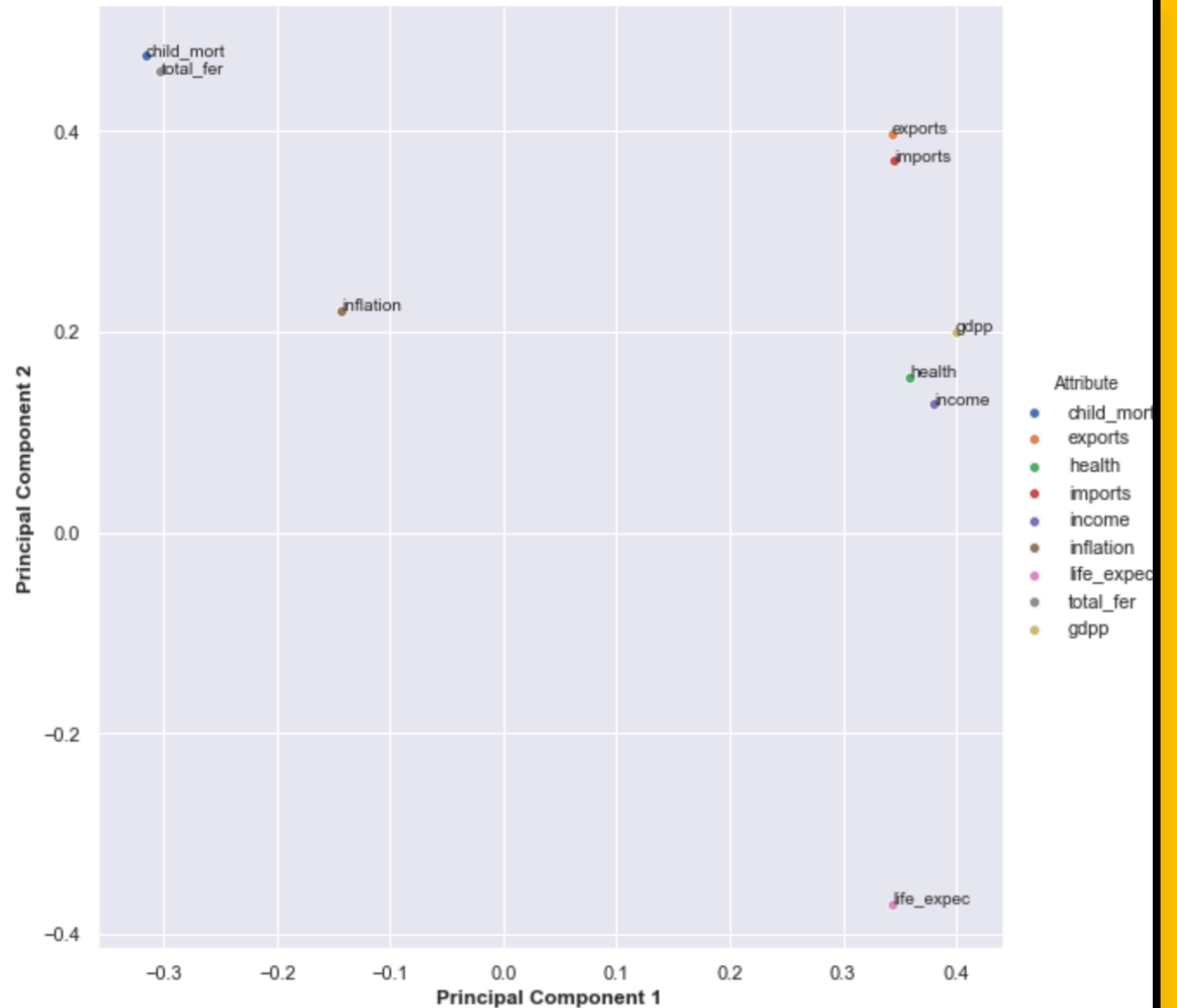
Scree Plot

- It is evident from the above Scree plot that more than 90% variance is explained by the first 3 principal components.

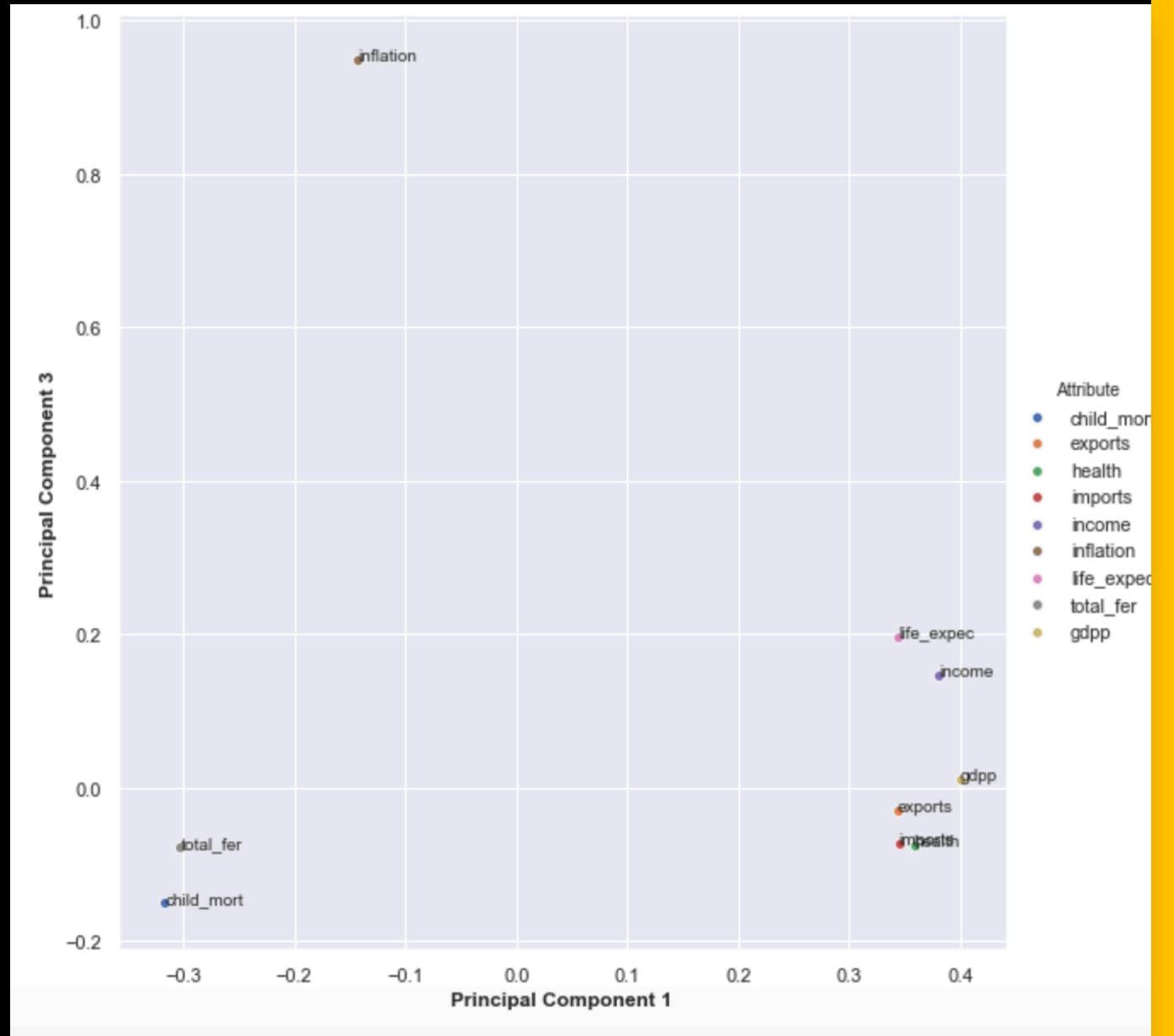


PCA

- life expectancy, income, gdp and health are very well explained by PC1.
- imports and exports are well explained by both the components PC1 and PC2.
- child mortality and total fertility are well explained by PC2.
- inflation is neither explained by PC1 nor with PC2



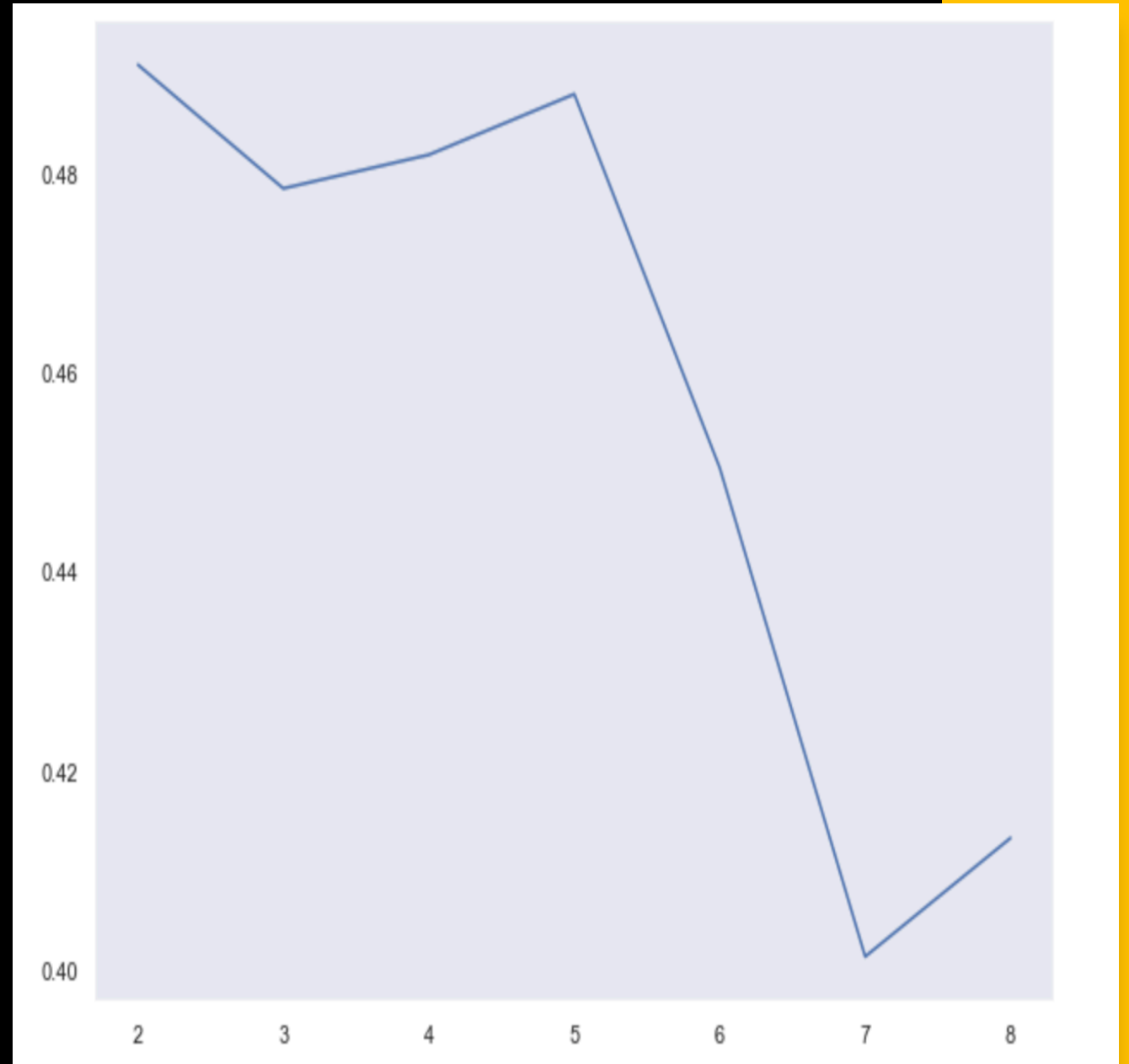
- inflation is well explained by PC3.



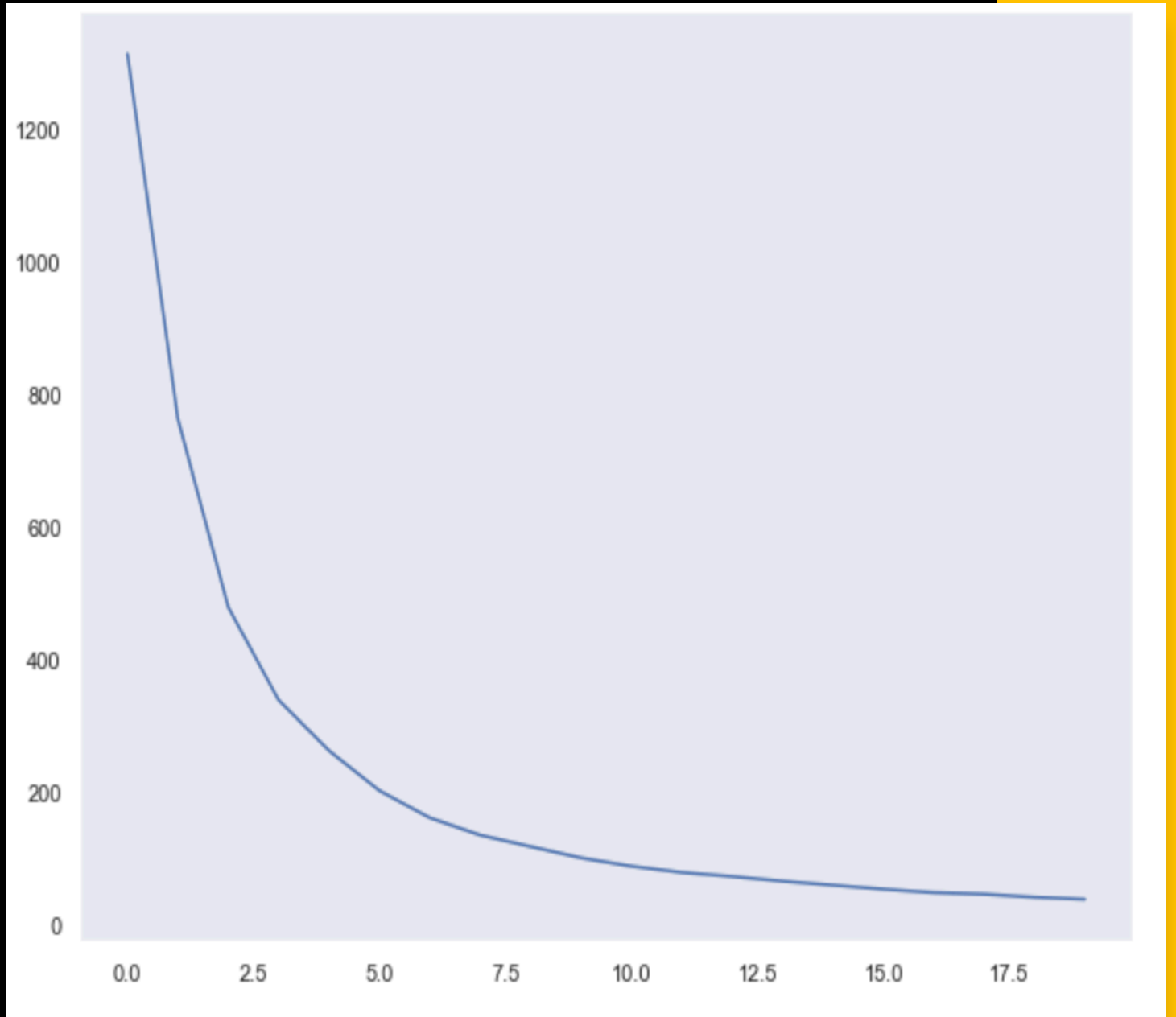
Clustering Process

- Both K-Means and Hierarchical Clustering are used on the 3 PCA Components.
- For both, K-Means and Hierarchical Clustering algorithms, number of clusters =5.
- The value of Hopkins Statistics after performing process >0.77

Silhouette Analysis

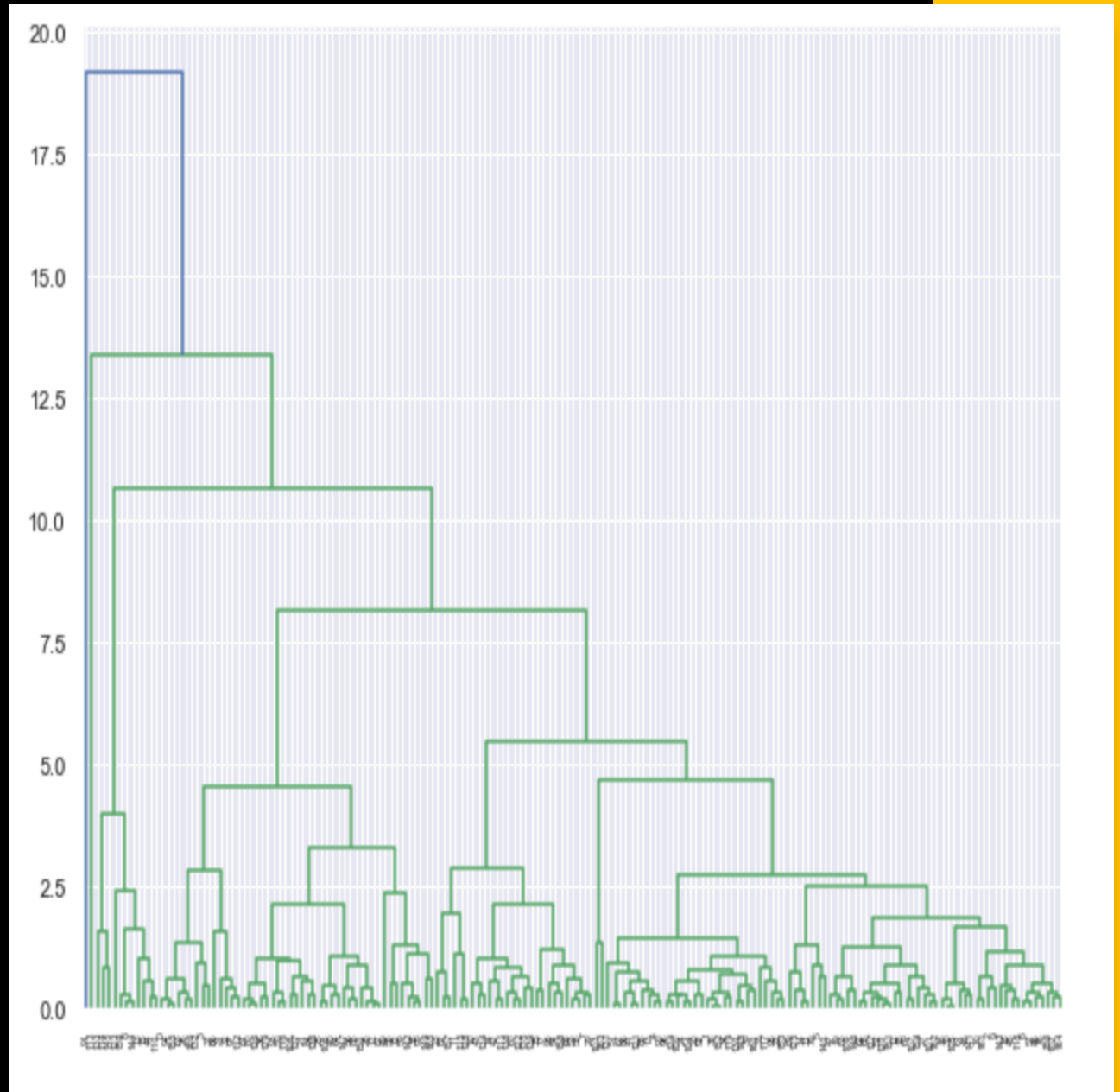


Sum of Squared Distances



Hierarchical Clustering

With Complete linkage



Conclusion

- The countries that require most help:

Afghanistan, Angola, Benin, Botswana, Burkina Faso, Burundi, Cameroon, Central African Republic, Chad, Comoros, Congo, Dem. Rep., Congo, Rep., Cote d'Ivoire, Equatorial Guinea, Eritrea, Gabon, Gambia, Ghana, Guinea, Guinea-Bissau, Haiti, Iraq, Kenya, Kiribati, Lao, Lesotho, Liberia, Madagascar, Malawi, Mali, Mauritania, Mozambique, Namibia, Niger, Pakistan, Rwanda, Senegal, Sierra Leone, Solomon Islands, South Africa, Sudan, Tanzania, Timor-Leste, Togo, Uganda, Yemen, Zambia, Nigeria

- These countries have very low net income per person, GDP Per Capita and high child mortality rate.

Factors Considered

- PCA was used above to reduce the variables involved and then clustering was done of countries based on those Principal components
- Later it was identified that few factors like child mortality, income etc plays a vital role in deciding the development status of the country and we built clusters of countries based on that.
- Based on those clusters it was identified that the below list of countries are in dire need of aid.
- The list of countries are subject to change as it is based on the few factors like Number of components chosen, Number of Clusters chosen, Clustering method used etc.which we have used to build the model.

Thank You