Optimizing Fund Raise

Which countries need the most help?

The problem

Company

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.

Context

We have raised 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.

Problem statement

How to optimally categorize and countries using some socio-economic and health factors that determine the overall development of the country and suggest the countries which the CEO needs to focus on the most.

Challenges deep-dive

Data Prep

Prepare the data

Reading the data

Check for outliers

Check for missing values

Understanding the data

Model building

RFM

- Created RFM - Recency, Frequency and monetary
- Outlier treatment and Rescaling

Clustering & PCA

Clusters and component analysis

Using the algorithms to create clusters

Validating the cluster

Identifying the countries that need help

Solution

Countries accurately identified

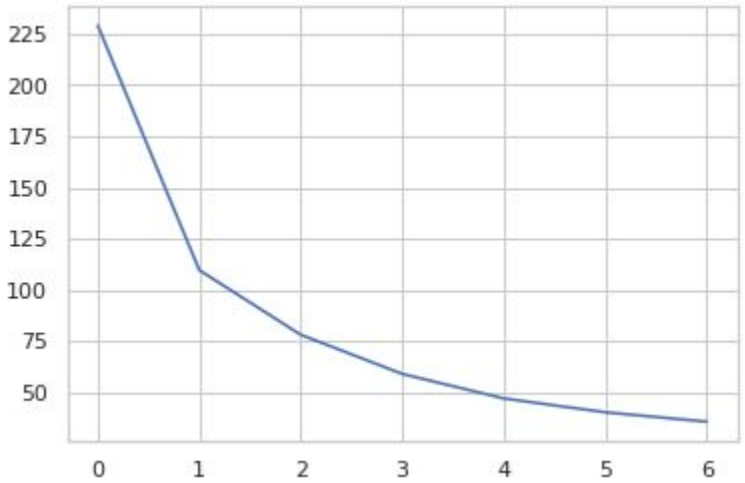
“The goal is to turn data into information, and information into insight.” - Clary Florin former CEO of Hewlett Packard

Implementation

Understanding

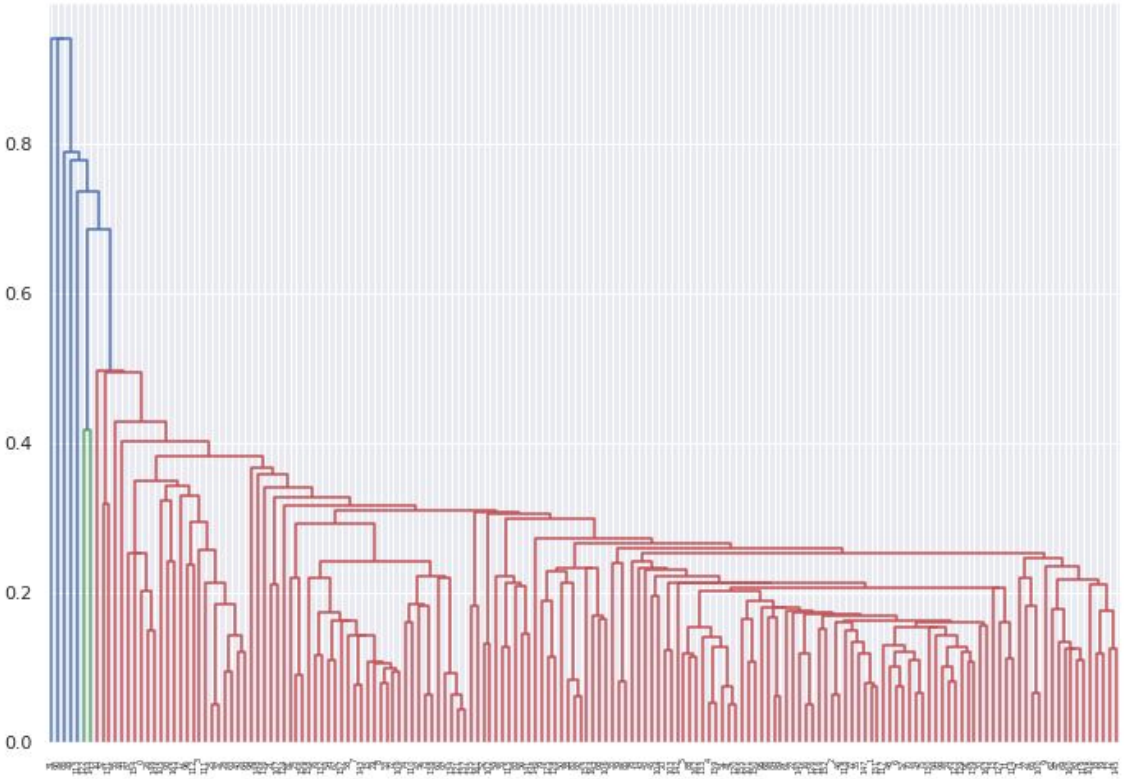
1. PCA is a linear transformation method and works well in tandem with linear models such as linear regression, logistic regression etc., though it can be used for computational efficiency with non-linear models as well
2. In the K-Means algorithm, you divided the data in the first step itself. In the subsequent steps, you refined our clusters to get the most optimal grouping. In hierarchical clustering, the data is not partitioned into a particular cluster in a single step. Instead, a series of partitions/merges take place, which may run from a single cluster containing all objects to n clusters that each contain a single object or vice-versa.

Silhouette Analysis



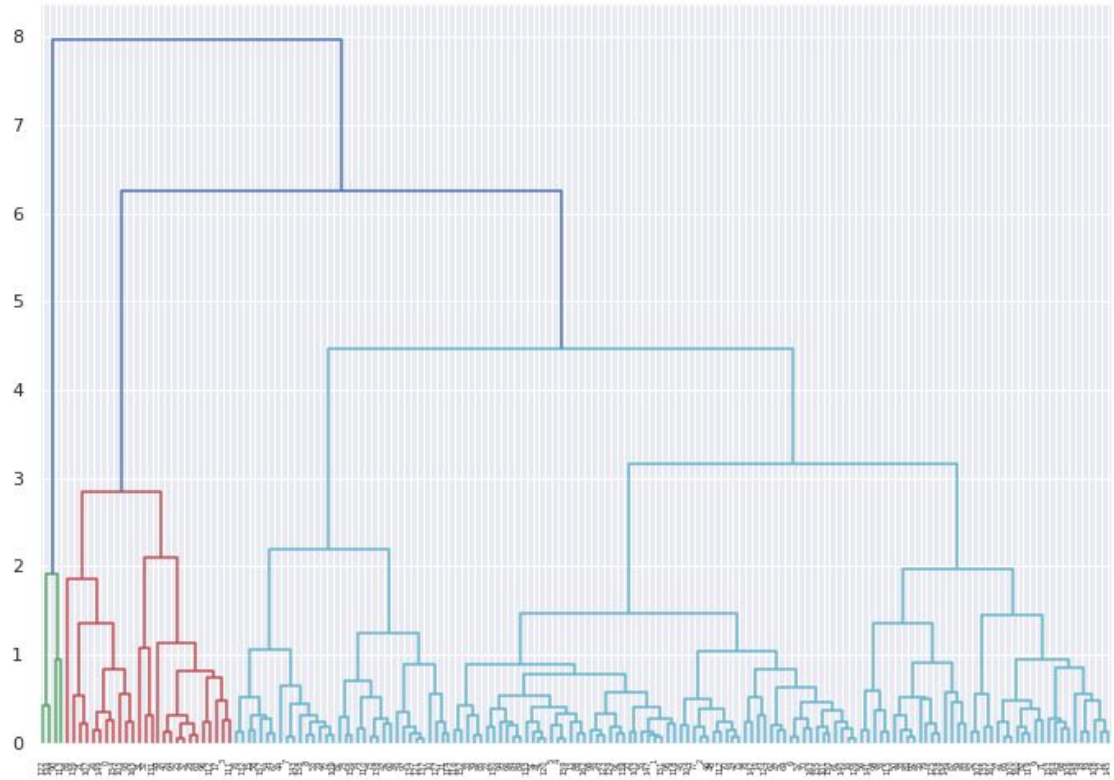
- The value of the silhouette score range lies between -1 to 1.
- A score closer to 1 indicates that the data point is very similar to other data points in the cluster,
- A score closer to -1 indicates that the data point is not similar to the data points in its cluster.

Single Linkage



This did create the linkage but could not derive the number of clusters needed for our problem statement. Linkages are not as structured as complete

Clusters created using Hierarchial Clustering



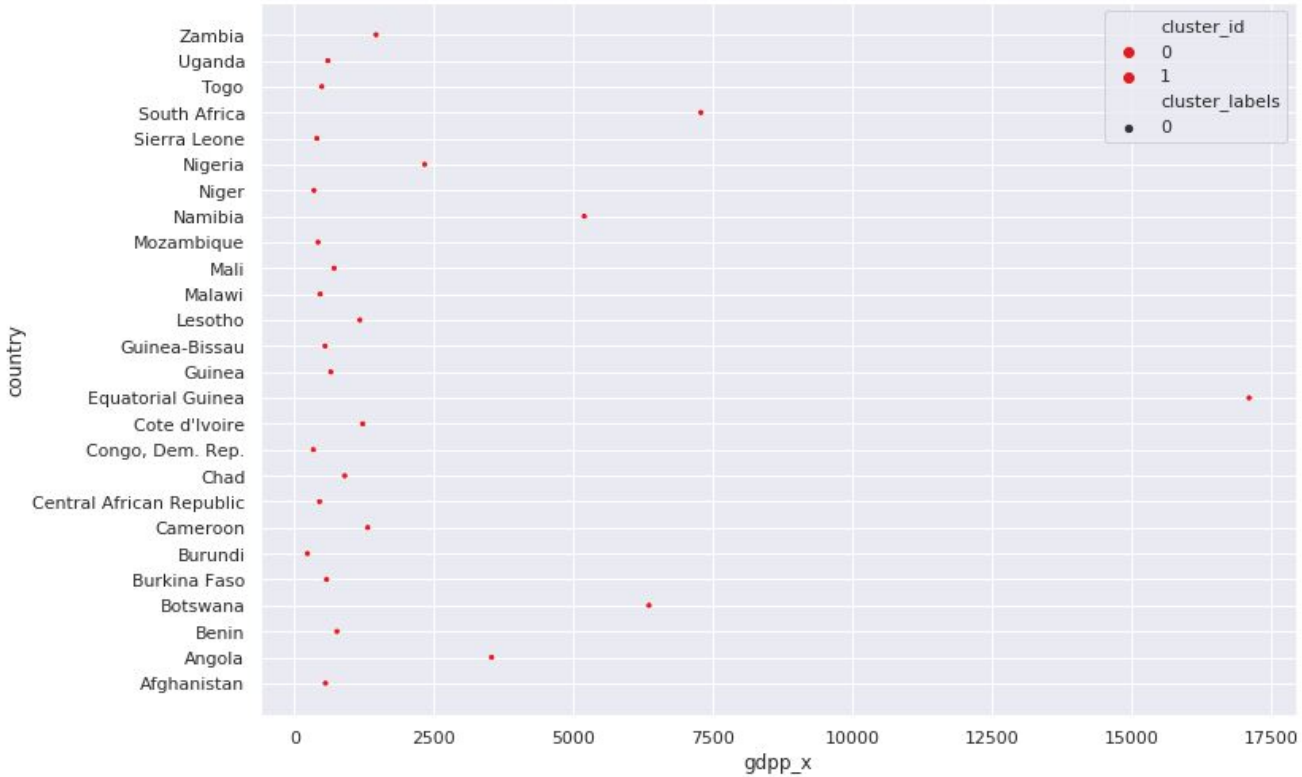
Now this is a complete cluster and we could cut at 3 clusters as you can see from the dendrogram.

Hence we could conclude in PCA we can directly use the PCA function, we could perform fit and transform on the RFM data, using which we could verify the model. In the final calculation also it was verified for the 3 clusters.
PCA doesnt change the total variance of the dataset, it only rearranges them in the direction of maximum variances.
So plotted a correlation matrix, there were no correlations between any two components.

Woohooo - to PCA!

We have effectively removed multicollinearity from our situation and our models will be stable.

Countries that need aid are



Thank you