UpGrad

# Clustering & PCA Assignment

By

Vyom Bhatt

# The Problem Statement

Health International, an International Humanitarian NGO, is committed to fighting poverty and providing the people of backward countries with basic amenities during times of distress. 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 their recent funding programmes, they've recently raised around $10million. Their main point of concern is identifying the countries that are in the most dire need of aid so that they can efficiently and strategically distribute the amount.

It is here where advanced analysis has been done to identify a list of countries that are in the most dire need of aid based on some socio-economic and health factors that determine the overall health of the country.

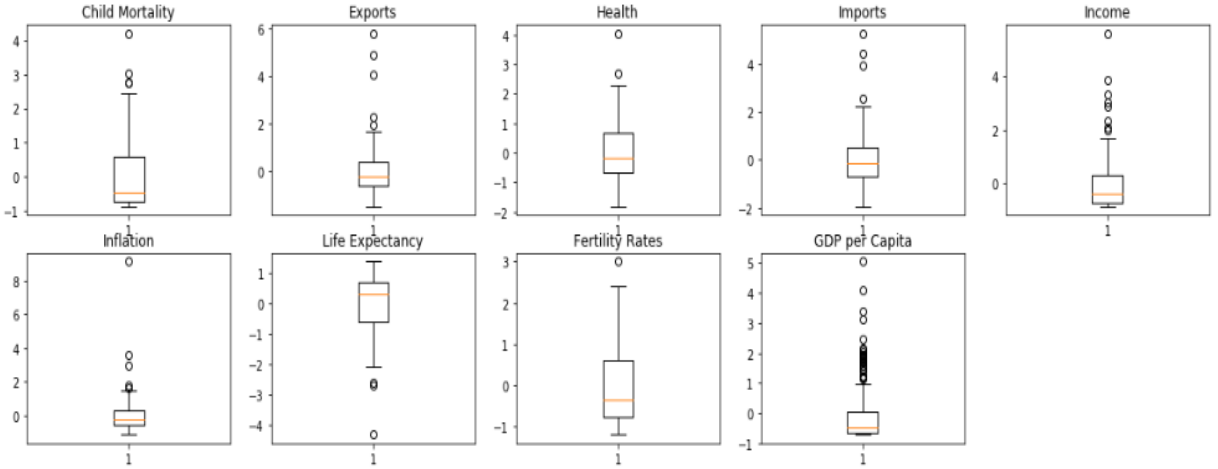
# Methodology

1. Initially performing EDA/Missing values check on the different variables provided in the dataset to visualize the data/countries
2. Outlier Analysis - considering we're looking for countries that are in the most dire need of aid, it doesn't make sense to get rid of the bottom outliers. Hence only the top outliers (above $95^{th}$ quantile) were treated
3. Data standardisation - to ensure all variables are on the same scale
4. Checking for multicollinearity
5. PCA – performing PCA on the dataset to treat multicollinearity and dimensionality reduction.
6. The principal components were visualized and analysed
7. Based on scree plot, decided on the number of components that explain at least 90% of the variance in the data
8. Principal component transformation was done on the original dataset to transform the original variables
9. Calculated the Hopkin's Statistic (to check if the data was fit for clustering),
10. Performed Silhouette Analysis and Elbow Curve analysis to identify the optimal number of clusters
11. K-means clustering was done on the data to with 4 clusters being created
12. Merged the principal component based data (with cluster info) with the original dataset and calculated statistics of the new clusters based on the original variables
13. Used binning concept to identify a set of countries that are most in need of financial aid
14. Repeated steps 11, 12, 13 but this time using Hierarchical Clustering to identify the list of countries

# Outlier Analysis

**Outlier Analysis**

As mentioned previously, it was important to deal with the upper outliers and **not** the bottom outliers as that is still our area of interest. Looking at the distribution of the different features provided:



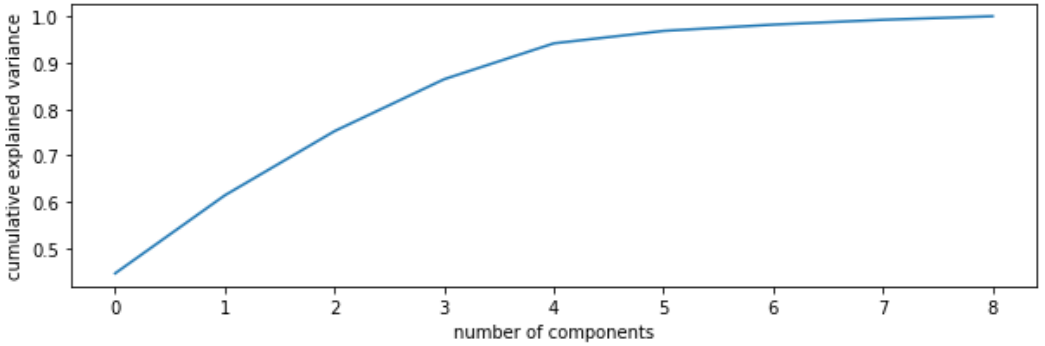

Considering we're only treating the outliers that exist for the really well performing countries, each feature would be treated differently based on whether the feature value is good or bad. Therefore, we'll be treating the **upper outliers for Exports, Health, Income, Life Expectance, GDP per Capita** and **lower outliers for Child Mortality, Imports, Inflation, Total Fertility.**

These outlier countries are those that are performing really well in terms of both health and socio-economic factors, and they included: Belgium, Iceland, Ireland, Malta, Netherlands, Slovak Republic, Slovenia and Switzerland. All of these outlier countries were **removed** from the dataset and we proceeded with the analysis

# PCA

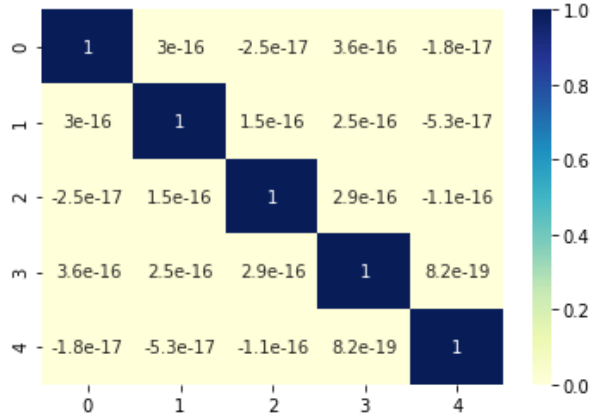
Post outlier treatment and standardisation, **PCA was performed** on the new dataset. PCA was performed with 2 major intentions: taking care of the multicollinearity in the data, and dimensionality reduction.

We used the **scree plot** to identify the number of principal components that explain at least 90% of variance in the data



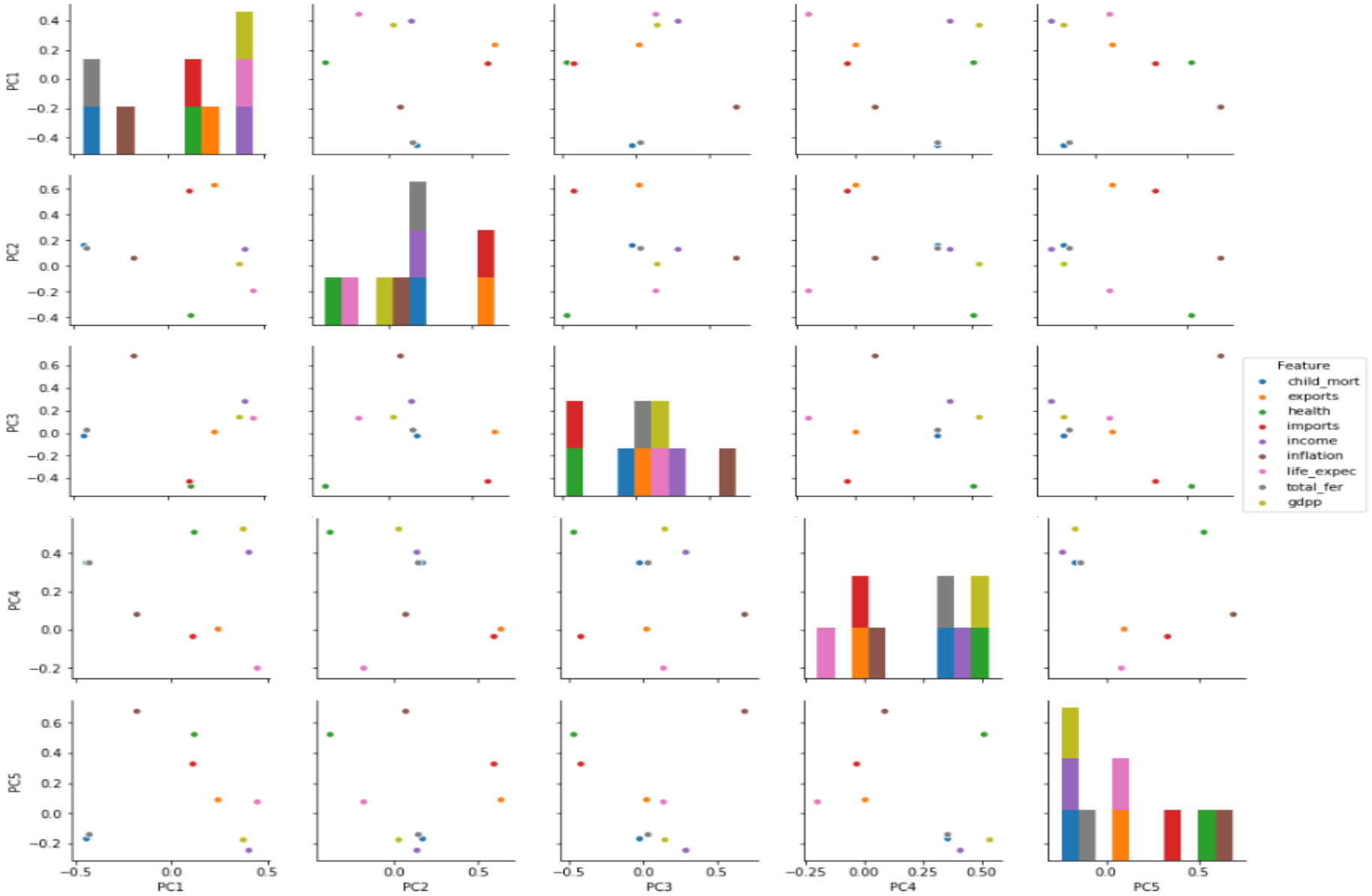

We can see that **5 principal components** explain over 90% of variance in the data

Now, let's see if we've taken care of the multicollinearity that existed in the data.




As we can see, there exists almost 0 correlation between any of the principal components

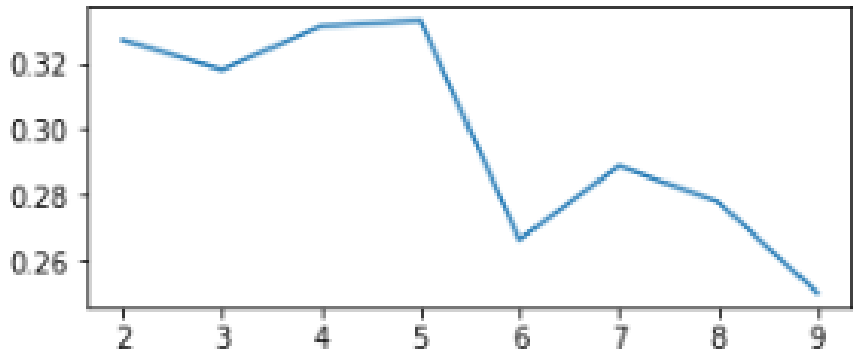
# Visualising the Principal Components



# K-Means Clustering

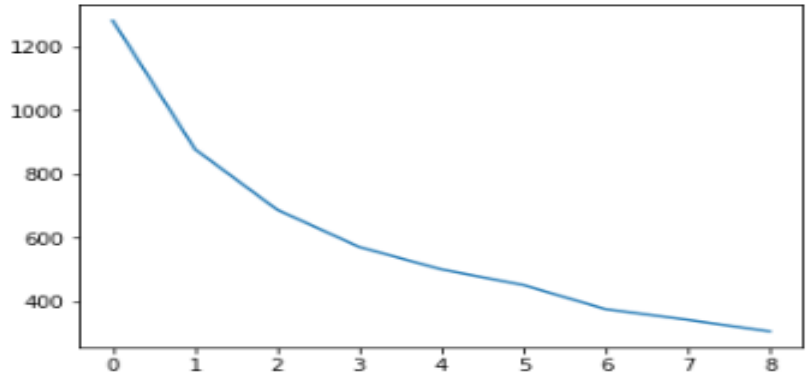
Next step was calculating the Hopkin's Statistic, which was **>0.7,** hence the data was good to be clustered.

On performing the Silhouette Analysis, we were able to identify the following trend:




We can see that creating 4 clusters would produce the optimal split of data

On checking the elbow curve for the above, we also observe that there is a decent drop in the y-axis at the point of x=4, which is a good indication that we can go ahead with making 4 clusters



# K-Means: Visualising the Clusters

Considering we have 5 different principal components, visualising the cluster together would have to be done on a 5-Dimensional space. This is why we've shown a pair-plot to compare every principal component against each other

# K-Means: Analysing the Clusters

The next step would be **merging** the PCA data with the clusters back to the initial dataset with the original features.

On analysing the new clusters based on the original features using the means, we see the following trend:

| ClusterID | Child Mortality | Exports | GDP per Capita | Health | Imports | Income | Inflation | Life Expectancy | Total Fertility |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.096151008 | -0.045588798 | -0.046130882 | 0.037561382 | 0.031257751 | -0.086850108 | -0.126184129 | -0.106440955 | 0.093674331 |
| 2 | 0.028967088 | -0.047657339 | -0.086909222 | -0.226167066 | -0.0731816 | -0.011368243 | 0.065413237 | -0.004318445 | -0.024751663 |
| 3 | 0.018719567 | -0.174551569 | -0.052903667 | 0.192698468 | -0.125578959 | -0.123876798 | -0.037008636 | -0.077946163 | 0.103457751 |
| 4 | 0.352846942 | -0.24656988 | -0.171897345 | 0.447536987 | 0.657168738 | 0.306322785 | 0.336133497 | -0.292998792 | 0.688336109 |

* The conditional formatting done based on **absolute scaled numbers** to indicate importance of each feature by cluster

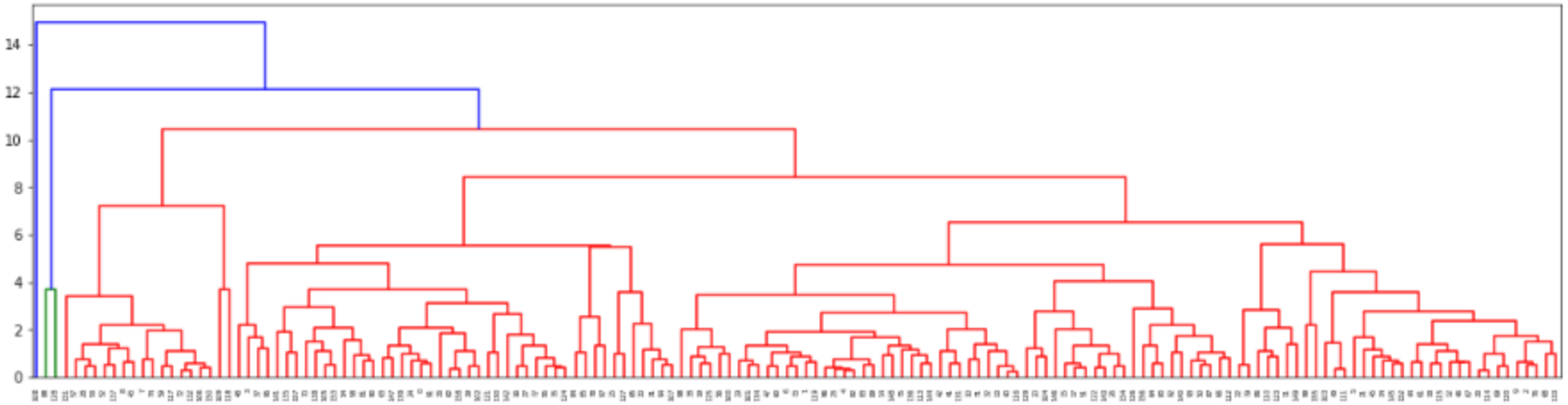
- We observe that all the features appear to be the highest in the 4$^{th}$ Cluster. The 1$^{st}$ and 2$^{nd}$ clusters include countries that are not so well off, and hence these are the clusters that include countries that are in dire need of financial aid.

**Binning**

In order to identify the final list of countries, we performed binning based on each of the variables using the means that have been calculated shown in the table above. The countries were identified such that their **child mortality, imports, total fertility is above the highest mean** and **exports, health, income, inflation, life expectancy and GDPP are below the lowest mean**

# Hierarchical Clustering

Hierarchical clustering was done on the 5 principal components identified in step 8 of our methodology. The corresponding **dendrogram** based on the 'complete' method of hierarchical clustering is as below




Here, we cut the dendrogram at 4 clusters based on our scree plot that was identified before doing the K-means plot in our analysis.

# Hierarchical: Analysing the Clusters

The next step would be **merging** the PCA data with the clusters back to the initial dataset with the original features.

On analysing the new clusters based on the original features using the means, we see the following trend:

| ClusterID | Child Mortality | Exports | GDP per Capita | Health | Imports | Income | Inflation | Life Expectancy | Total Fertility |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.030707525 | -0.05647381 | -0.076190385 | -0.13290675 | -0.05413051 | -0.03641225 | 0.015842622 | -0.015776599 | 0.00495259 |
| 2 | 0.144146479 | -0.16432693 | -0.000566211 | 0.291600084 | -0.06109579 | -0.1406215 | -0.15566598 | -0.194921049 | 0.195845595 |
| 3 | 0.352846942 | -0.24656988 | -0.171897345 | 0.447536987 | 0.657168738 | 0.306322785 | 0.336133497 | -0.292998792 | 0.688336109 |
| 4 | 0.439633273 | 0.244090963 | -0.424151939 | -0.01299262 | 0.570426249 | -0.4504958 | -0.3993899 | -1.344367201 | 0.430714395 |

* The conditional formatting done based on **absolute scaled numbers** to indicate importance of each feature by cluster

- We observe that a large majority of the features appear to be the highest in the 4$^{th}$ Cluster. The 1$^{st}$ and 2$^{nd}$ clusters include countries that are not so well off, very similar to the output of the K-means based clusters

**Binning**

In order to identify the final list of countries, we performed binning based on each of the variables using the means that have been calculated shown in the table above (similar to the process followed for K-means clustering). The countries were identified such that their **child mortality, imports, total fertility is above the highest mean** and **exports, health, income, inflation, life expectancy and GDPP are below the lowest mean**

# Final Results

Both the K-means and Hierarchical types of clustering returned results very identical to each other. After the process of binning, the **countries identified as those that are in the direst need of financial aid** from Help International include:

Niger

Central African Republic

Comoros

Gambia

Benin

Mali

Senegal

Kenya

Cameroon