




Clustering Assignment

HELP International NGO

Submitted By:
Srinivasaragavan V



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 programmes, 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 & Goals

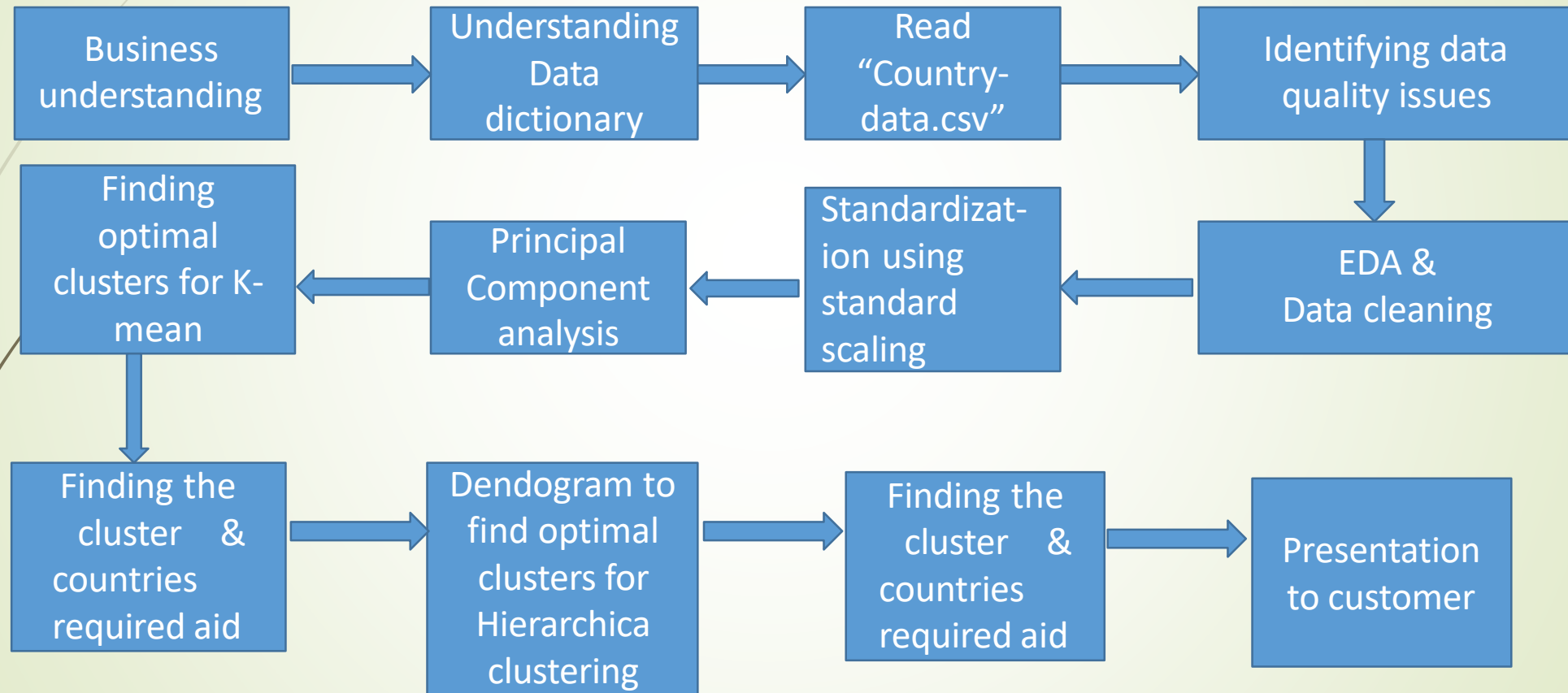
Business Objectives:

- The objective is to categorise the countries using some socio-economic and health factors that determine the overall development of the country.
- Suggest the countries which the CEO needs to focus on the most for funding purposes.

Goals of Analysis:

- Perform PCA on the dataset and obtain the new dataset with the Principal Components.
- Perform K-means and Hierarchical clustering on this dataset and create clusters.
- The final list of countries depends on the number of components that you choose and the number of clusters that you finally form.

Problem solving methodology



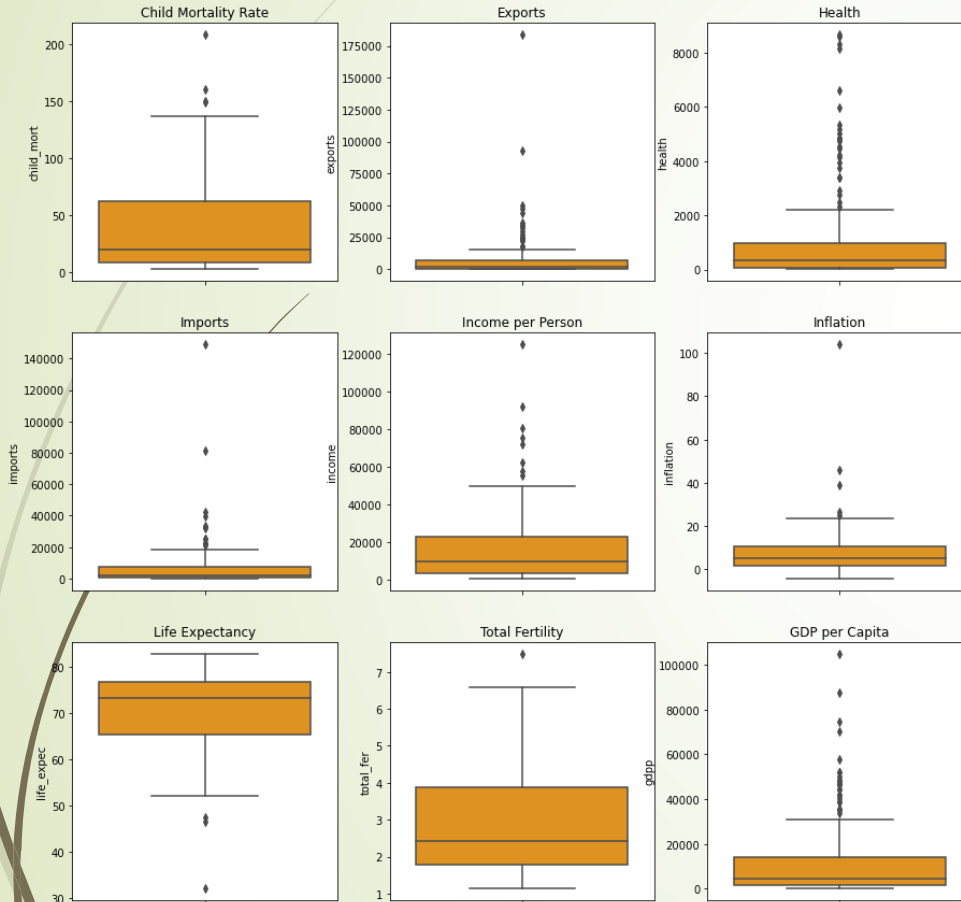


Exploratory Data Analysis



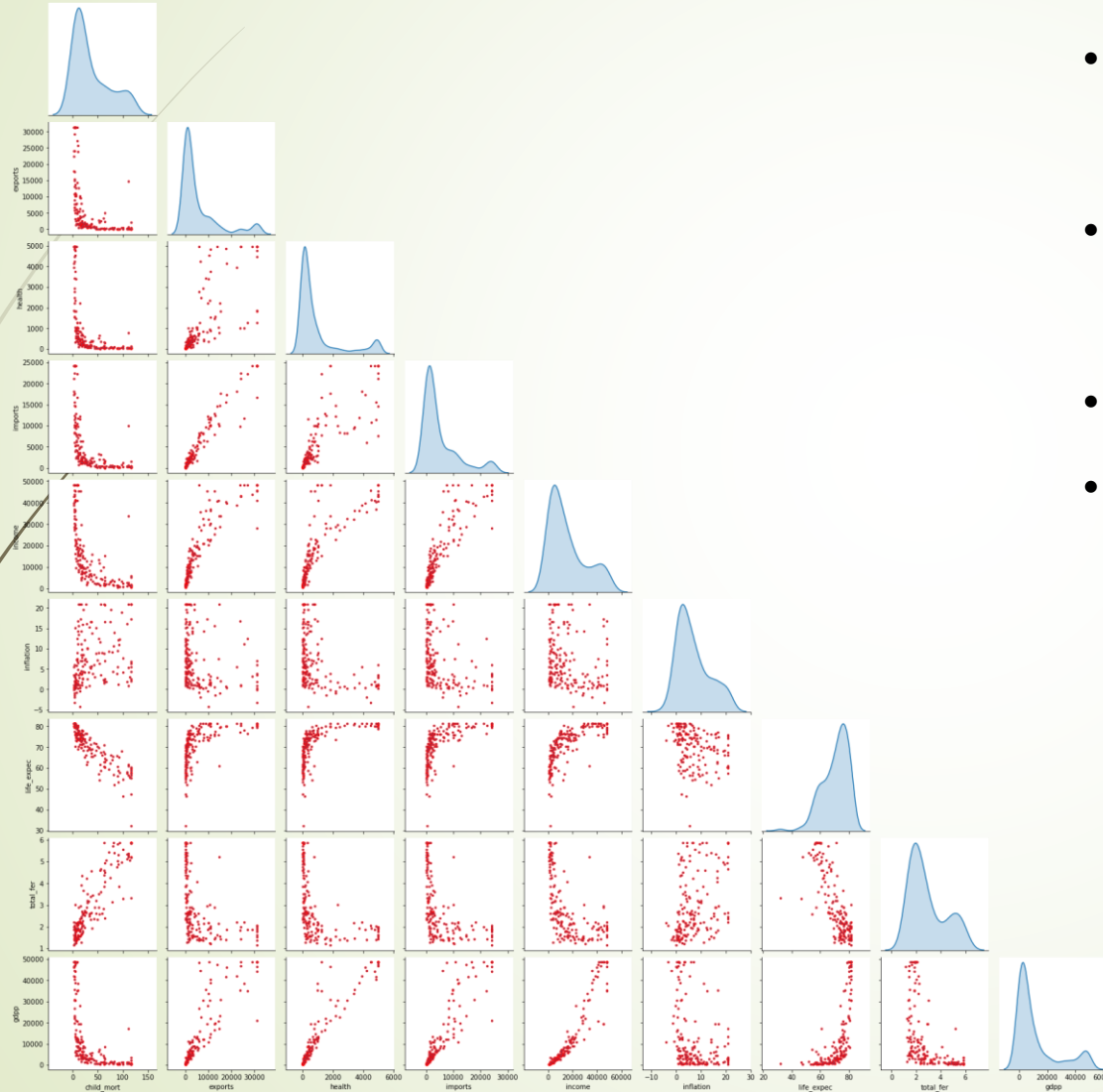
- This analysis is done to identify the 5 underdeveloped countries which are in need of the aid at most based on the following factors
 - low life_expectency, low gdpp, low income, low imports & exports, high inflation,high child_mortality, high total_fer
-
- Univariate – Outlier Analysis
 - Bi variate Analysis
 - Correlation analysis

Univariate – (Outliers Analysis)



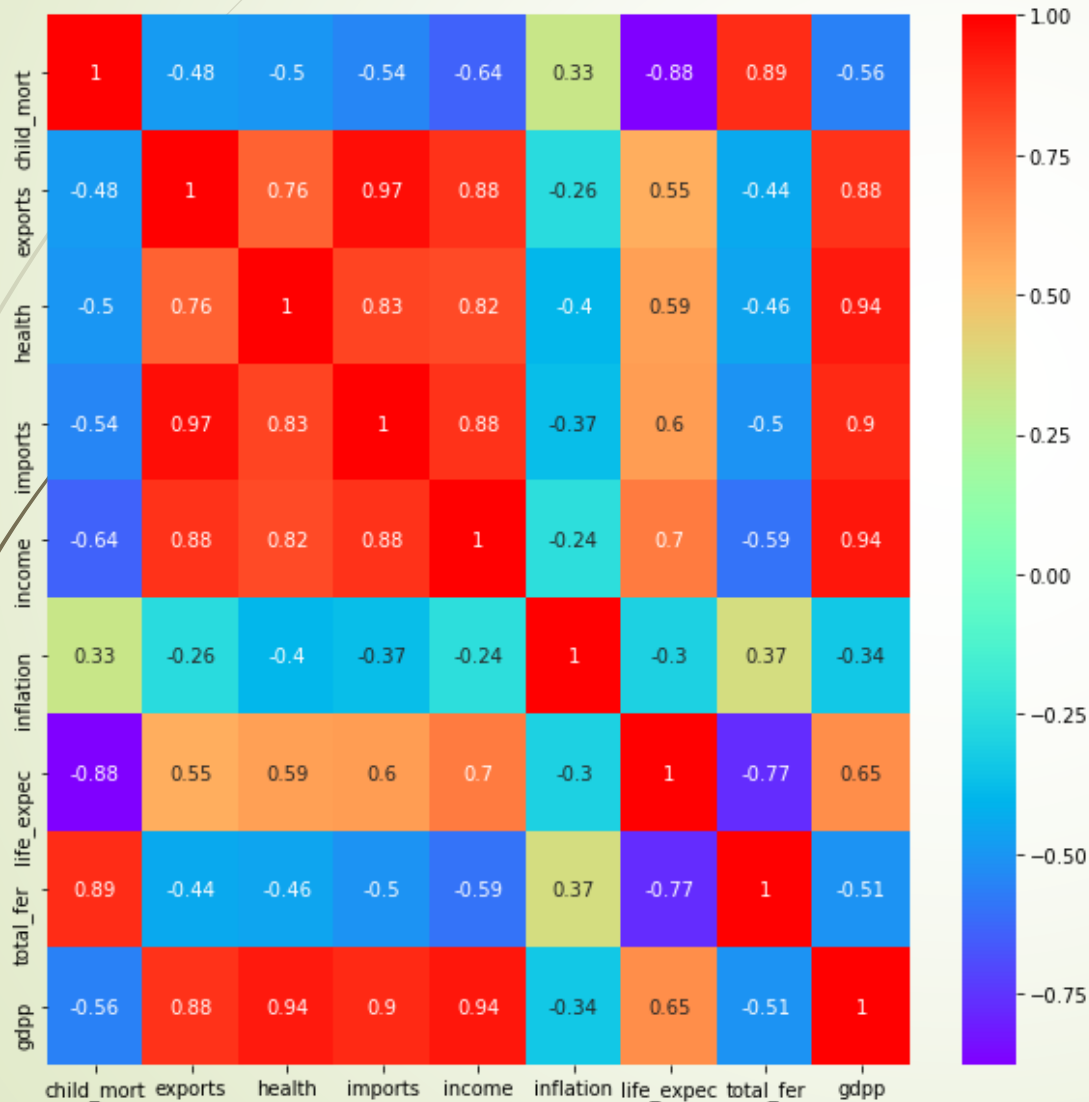
- There are a number of outliers in the data.
- Since, the K-Means algorithm tries to allocate each of the data point to one of the clusters, outliers have serious impact on the performance of the algorithm and prevent optimal clustering.
- Keeping in mind we need to identify backward countries based on socio economic and health factors.
- There could be a possibility in child mort's subplot where those outlier-countries could be the needy ones because of the high child-mortality rate.
- We will cap the outliers to values accordingly for analysis carefully .

Bi Variate Analysis



- `gdpp` and `total_fer` seem to be a bimodal rest of the features seems to be an uni-modal
- Few features are left skewed and few features are right skewed
- Lot of variance is found between the features
- The distribution of features points out the behaviour of the data and their ranges. We must standardize the data so as to avoid discrepancies in evaluation.

Correlation Analysis



- Imports have high positive correlation with Exports
- Income has fairly high positive correlation with Exports
- Life Expectancy has fairly high positive correlation with Income
- Total Fertility has very high positive correlation with Child Mortality
- GDPP has very high positive correlation with Income
- GDPP has fairly high positive correlation with Life Expectancy
- Total Fertility has fairly high negative correlation with Life Expectancy



Standardization Of Data



- "Standardization of data, that is, converting the m in to z-scores with mean 0 and standard deviation 1, is important for 2 reasons in K-Means algorithm:
- Since we need to compute the Euclidean distance between the data points, it is important to ensure that the attributes with a larger range of values do not out-weight the attributes with smaller range.
- Thus, scaling down of all attributes to the same normal scale helps in this process.
- The different attributes will have the measures in different units. Thus, standardisation helps in making the attributes unit-free and uniform.



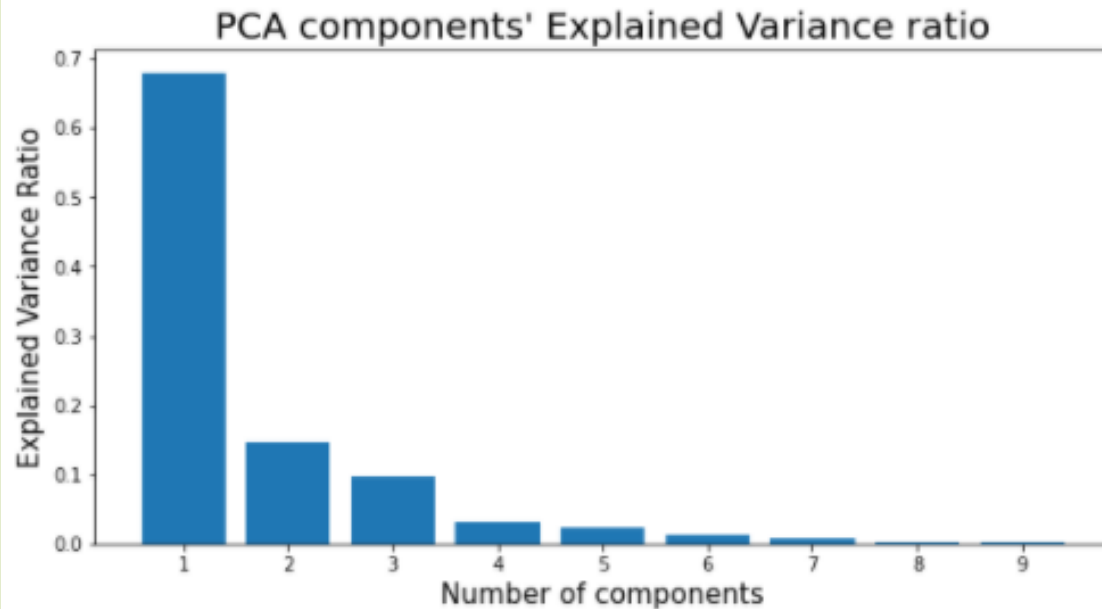
Principal Component Analysis



- Principal Component Analysis (PCA) is a popular technique for deriving a set of low dimensional features from a large set of variables. Sometimes reduced dimensional set of features can represent distinct no. of groups with similar characteristics.
- Principal component analysis (PCA) is one of the most commonly used dimensionality reduction techniques in the industry. By converting large data sets into smaller ones containing fewer variables, it helps in improving model performance, visualising complex data sets, and in many more areas.

Explained variance ratio

The cumulative variance of the first 5 principal components is 0.97599

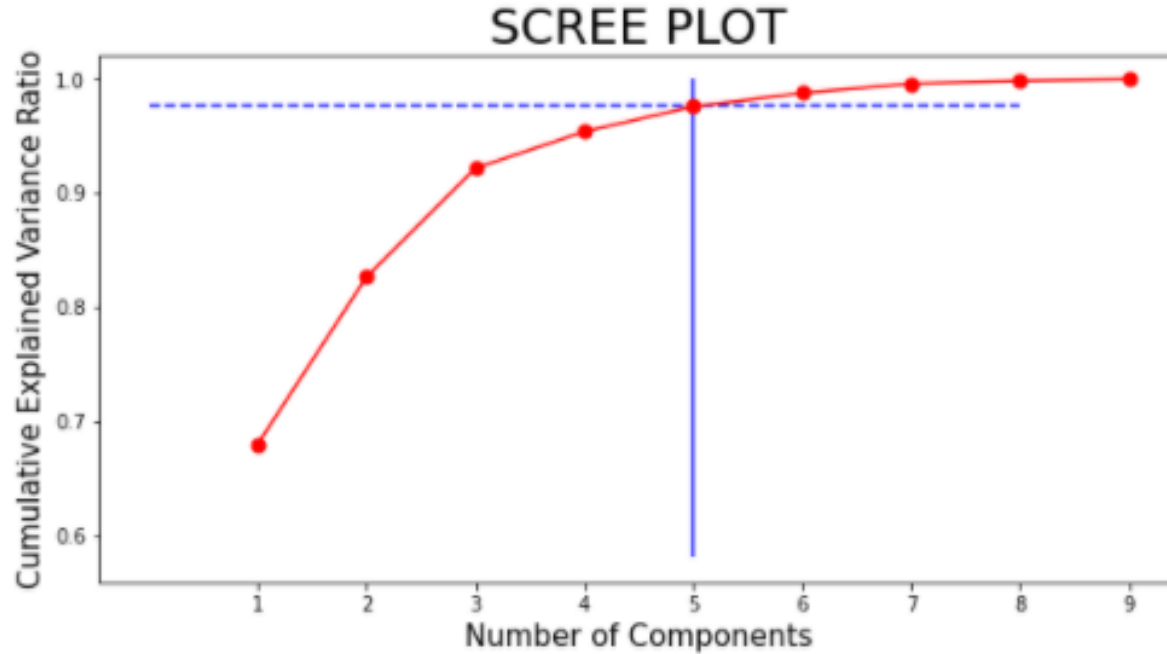


Inferences:

- First component variance explained is almost 65%.
- For second component variance explained is almost 20%.
- From the plot we could see that the variance drop off point occurs in between 1 and 2 component

- `explained_variance_` : array, shape (n_components,) The amount of variance explained by each of the selected components.
- `explained_variance_ratio_` : array, shape (n_components,) Percentage of variance explained by each of the selected components.

Scree Plot

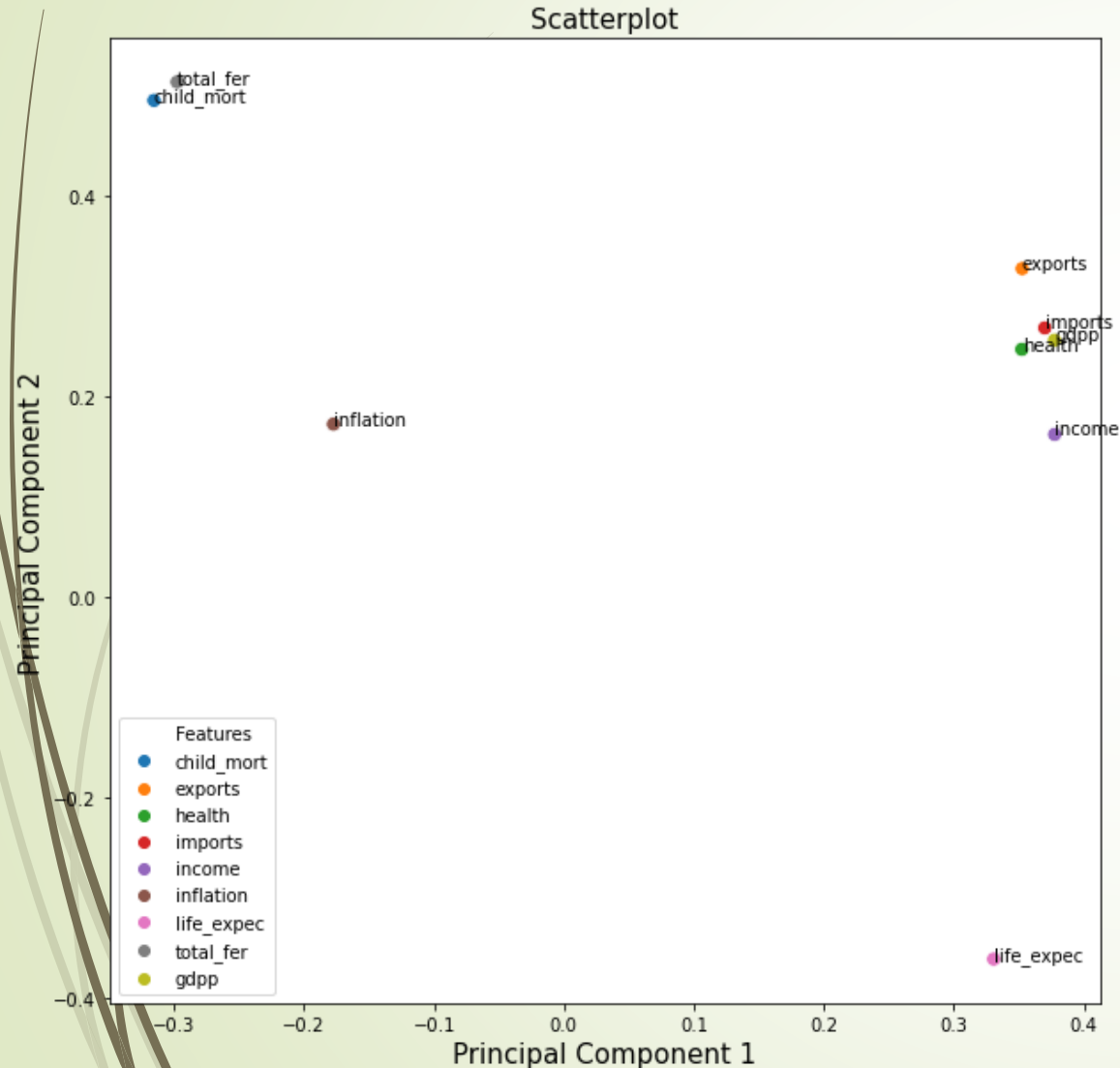


Inferences:

- From above scree plot, Around 97% of the information is being explained by 5 components.
- The plot is following an upward trend. After crossing 5, it is almost steady.

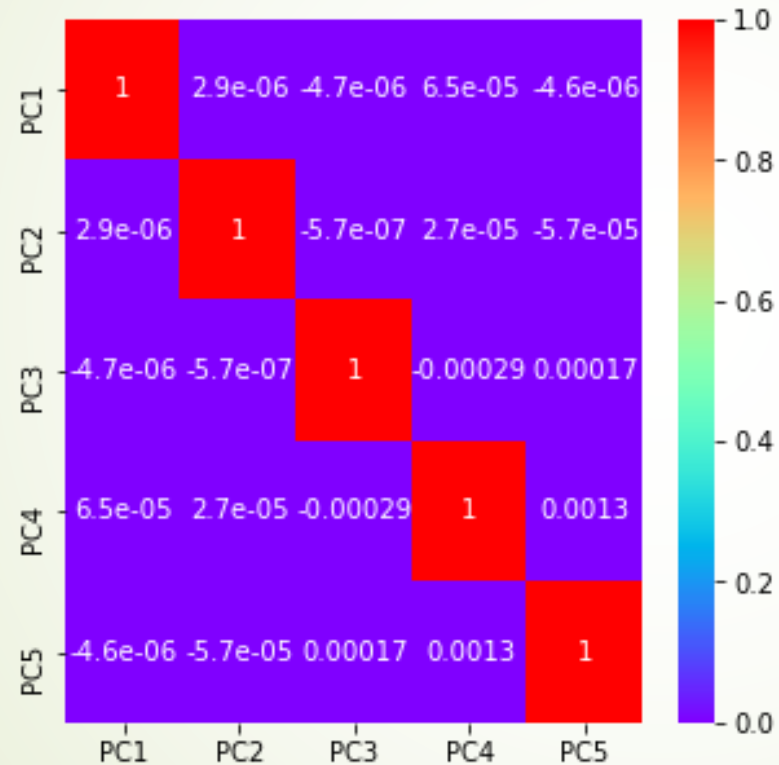
- A scree plot is a line plot of the eigenvalues of factors or principal components in an analysis.
- The scree plot is used to determine the number of factors to retain in an exploratory factor analysis (FA) or principal components to keep in a principal component analysis (PCA).
- A scree plot always displays the eigenvalues in a downward curve, ordering the eigenvalues from largest to smallest. According to the scree test, the "elbow" of the graph where the eigenvalues seem to level off is found and factors or components to the left of this point should be retained as significant.

Visualising 2 main Principal Components



- From the plot, We can see that 1st Principal Component (X-axis) is gravitated mainly towards features like: imports, exports, gdpp, income, health.
- 2nd Principal Component (Y-axis) is gravitated mainly towards features like: child mort , total_fer.
- If you recall, correlation between imports and exports was 0.97. Now we can surely confirm it by looking the above plot.
- Inflation is neither gravitated to 1st Principal Component nor to 2nd Principal Component

Incremental PCA Correlation Check



The correlation among the attributes is almost 0.

Hopkins Statistic

- The Hopkins statistic (introduced by Brian Hopkins and John Gordon Skellam) is a way of measuring the cluster tendency of a data set. It acts as a statistical hypothesis test where the null hypothesis is that the data is generated by a Poisson point process and are thus uniformly randomly distributed.
- A value close to 1 tends to indicate the data is highly clustered, random data will tend to result in values around 0.5, and uniformly distributed data will tend to result in values close to 0.

Hopkins score

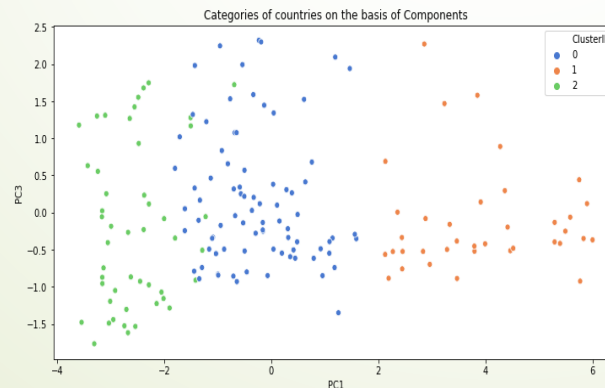
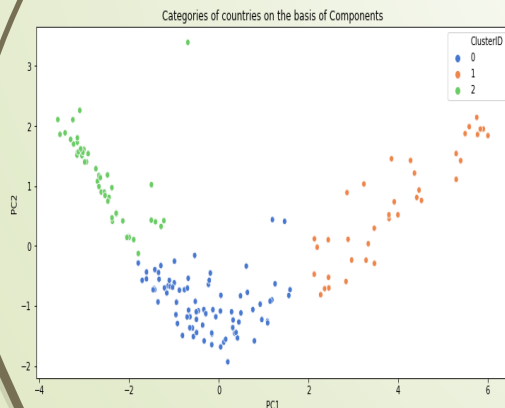
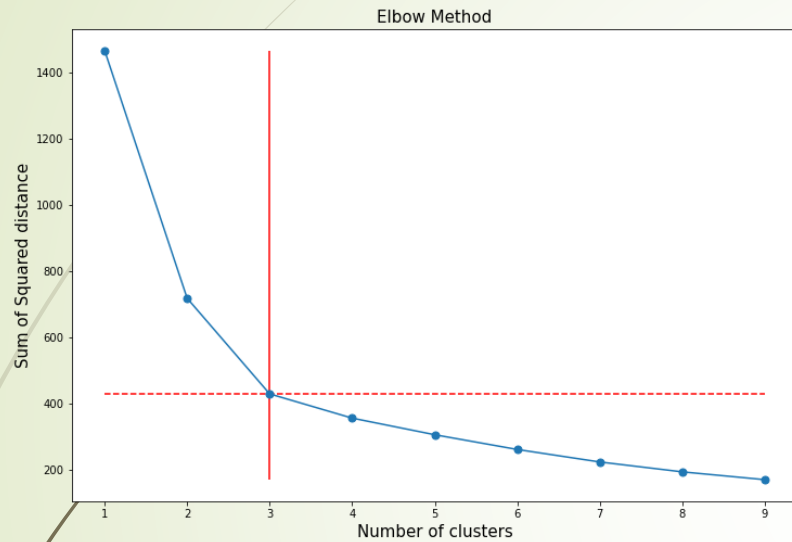
```
round(hopkins(pca_data),5)
```

0.81329

Inference:

Hopkins Statistic over .70 is a good score that indicated that the data is good for cluster analysis. - A 'Hopkins Statistic' value close to 1 tends to indicate the data is highly clustered, random data will tend to result in values around 0.5, and uniformly distributed data will tend to result in values close to 0.

K-means Clustering

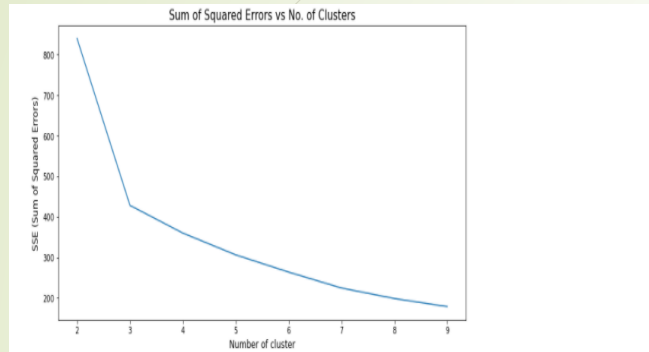


K-means algorithm explores for a preplanned number of clusters in an unlabelled multidimensional dataset, it concludes this via an easy interpretation of how an optimized cluster can be expressed.

Key features of k-means clustering;

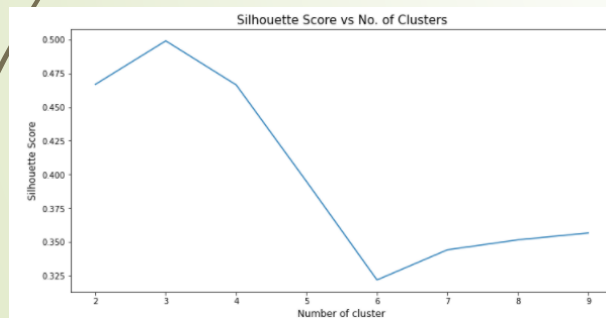
- It is very smooth in terms of interpretation and resolution.
- For a large number of variables present in the dataset, K-means operates quicker than Hierarchical clustering.
- While redetermining the cluster center, an instance can modify the cluster.
- k-means reforms compact clusters.
- It can work on unlabelled numerical data.

Silhouette Analysis



Inferences:

- We can see "knee" like bent at both 3 and 4, So considering no. of clusters = 3 seems a better choice. Still, we will analyse further to decide between 3 and 4.
- But, same was observed in the elbow curve method also which makes the approach certain

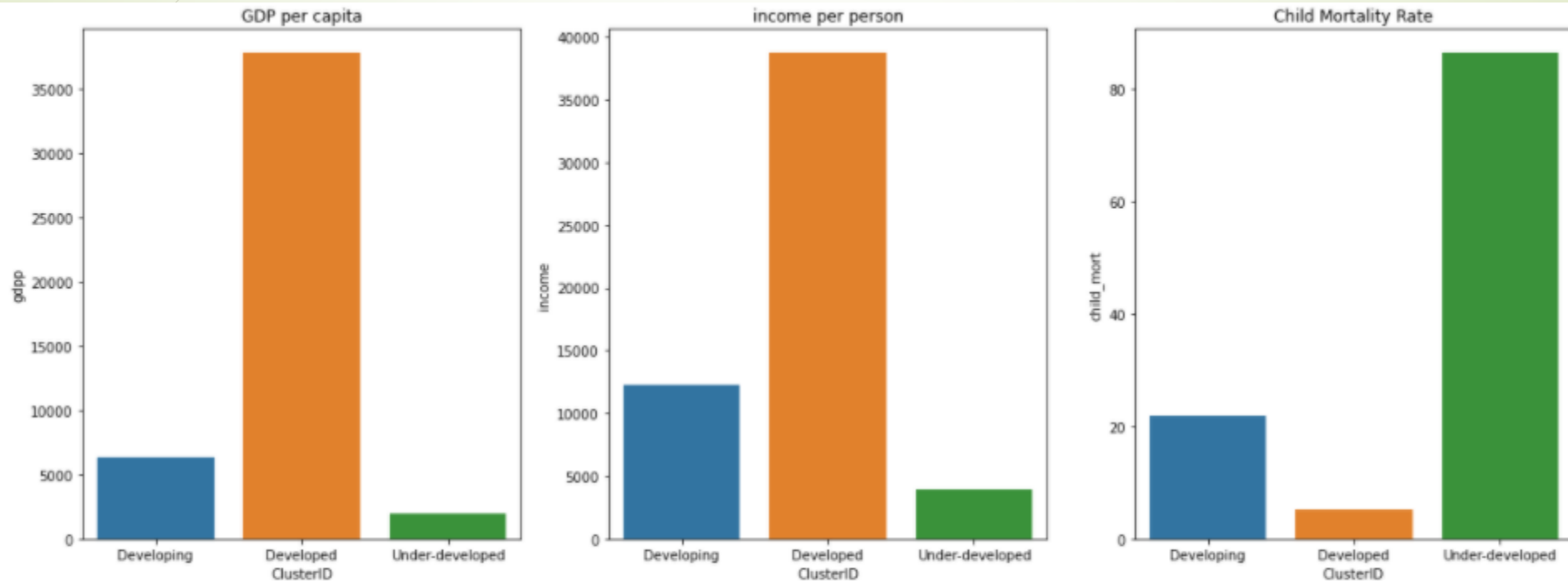


Inferences:

- No. of clusters = 3 seems the best choice here as well. As the highest peak is at 3

- To check how good is our K-Means clustering model. Silhouette Coefficient is one such metric to check that. The Silhouette Coefficient is calculated using:
- $$\text{silhouette score} = \frac{p - q}{\max(p, q)}$$
- p is the mean distance to the points in the nearest cluster that the data point is not a part of
- q is the mean intra-cluster distance to all the points in its own cluster.
- 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.

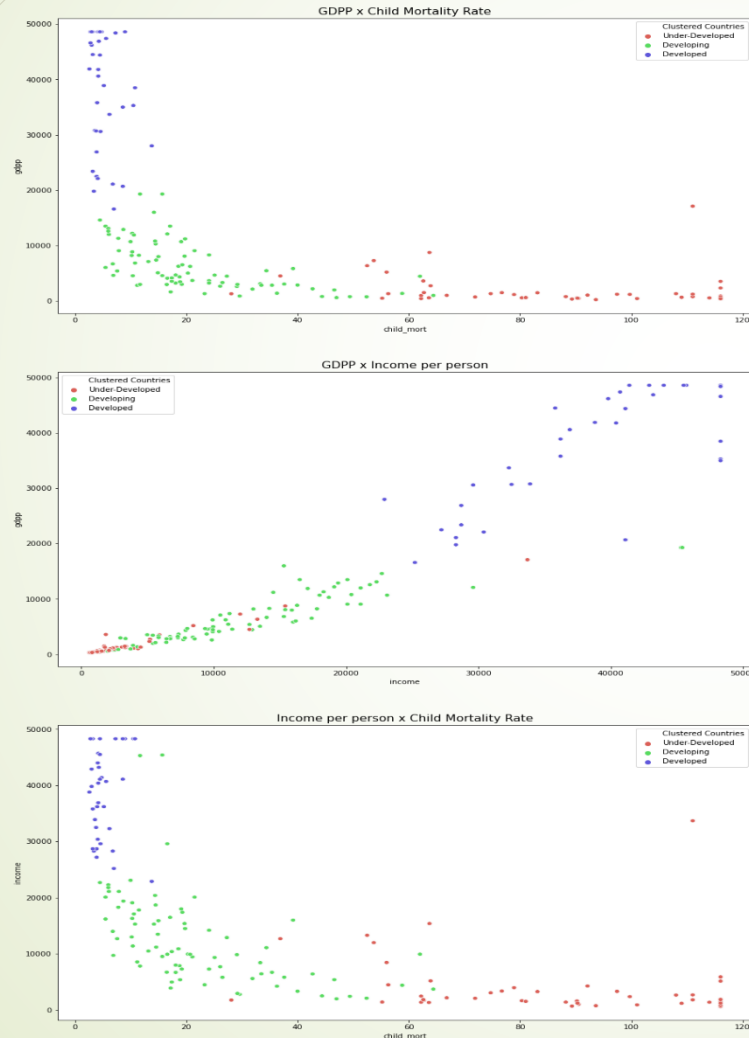
Cluster based comparisons



Inference:

- All the developed countries are having high GDP per capita values, developing countries are having average GDP per capita values and under developed countries are having the least GDP values.
- All the developed countries are having high income per person, developing countries are having average income per person and under developed countries are having the least income per person.
- All the developed countries are having low number of death of children under 5 years of age per 1000 live births, developing countries are having average death rate and under developed countries are having the high death rate.

Cluster based comparisons



- In `gdp x child_mort`, there is some clustering where `gdp` is more, there `child-mort` is low, which is a FACT.
- - In `gdp x income`, there is some clustering where `gdp` is more, then `income` is also more.
- - In `income x child_mort`, there is some clustering where if `child_mort` is more, then `income` is less.

Top Developed 10 Countries based on GFPP, Income, Child Mortality

Top 10 developed countries based on high GDPP

Australia
Denmark
Switzerland
Sweden
Qatar
Norway
Netherlands
Luxembourg
Ireland
United States

Top 10 developed countries based on high income

United States
United Arab Emirates
Switzerland
Brunei
Singapore
Qatar
Norway
Luxembourg
Kuwait
Ireland

Top 10 developed countries based on child low mortality

Iceland
Luxembourg
Singapore
Sweden
Finland
Japan
Slovenia
Norway
Czech Republic
Cyprus

Top 10 Developing Countries based on GFPP, Income, Child Mortality

Top 10 Developing countries based on high GDPP

Oman
Saudi Arabia
Barbados
Estonia
Venezuela
Croatia
Hungary
Chile
Poland
Antigua and Barbuda

Top 10 Developing countries based on high income

Saudi Arabia
Oman
Libya
Russia
Estonia
Hungary
Poland
Malaysia
Lithuania
Seychelles

Top 10 Developing countries based on child low mortality

Estonia
Croatia
Belarus
Poland
Hungary
Lithuania
Montenegro
Bosnia and Herzegovina
Serbia Latvia

Top 10 Under Developed Countries based on GFPP, Income, Child Mortality

Top 10 Under Developed countries based on high GDPP

Equatorial Guinea

Gabon

South Africa

Botswana

Namibia

Iraq

Timor-Leste

Angola

Congo,

Rep. Nigeria

Top 10 Under Developed countries based on high income

Equatorial Guinea

Gabon

Botswana

Iraq

South Africa

Namibia

Angola

Congo, Rep.

Nigeria

Yemen

Top 10 Under Developed countries based on child low mortality

Solomon Islands

Iraq

Botswana

South Africa

Eritrea

Namibia

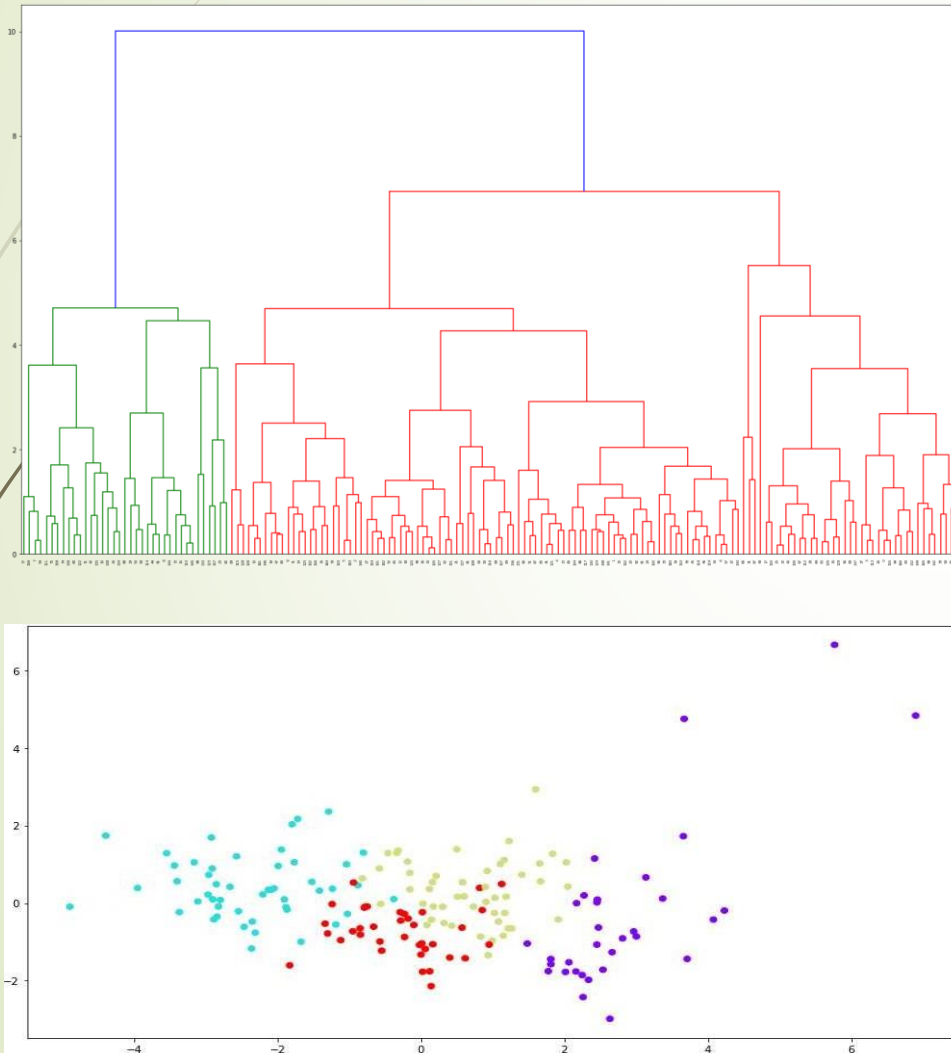
Yemen

Kenya

Madagascar

Timor-Leste

Conclusions- Hierarchical clustering



- The result of the hierarchical clustering algorithm is shown by a dendrogram, which starts with all the data points as separate clusters and indicates at what level of dissimilarity any two clusters were joined
- Once we obtain the dendrogram, the clusters can be obtained by cutting the dendrogram at an appropriate level. The number of vertical lines intersecting the cutting line represents the number of clusters.



Conclusions

Top 10 countries having dire need of aid based on overall conditions as per K-Means

Burundi
Congo, Dem. Rep.
Central African Republic
Guinea-Bissau
Afghanistan
Burkina Faso
Benin
Guinea
Comoros
Chad
Cote d'Ivoire
Cameroon