# Assignment

# “PCA and Clustering”

Dated: 19th August 2019

Submitted By:

Name: Vikram mathur

Roll Number: DDS1910599

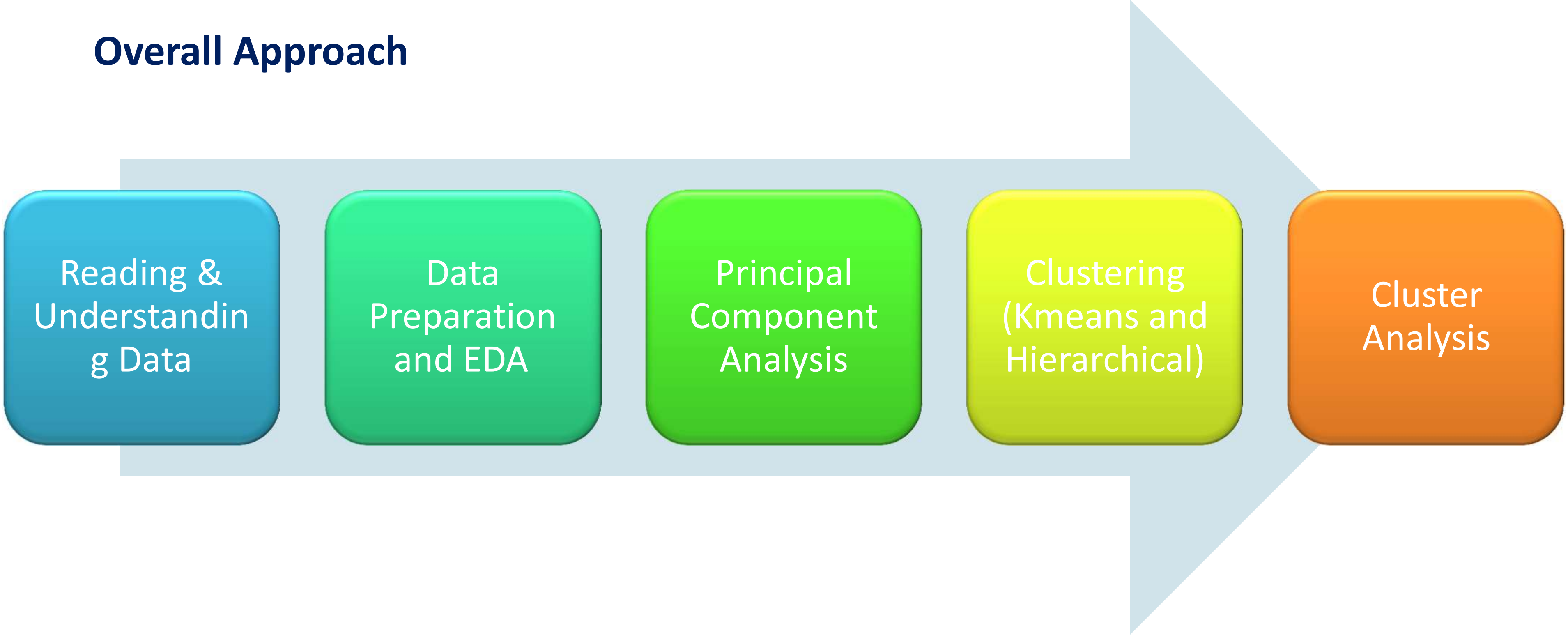
VikramMathur3012@gmail.com

# Problem Statement & Overall Approach

## Problem Statement

- Identifying at least 5 countries in direct need of aid for utilization of $ 10 million funding available with HELP International NGO.
- Objective is to determine the overall development of the countries based on socio-economic and health factors to take this decision.

## Overall Approach

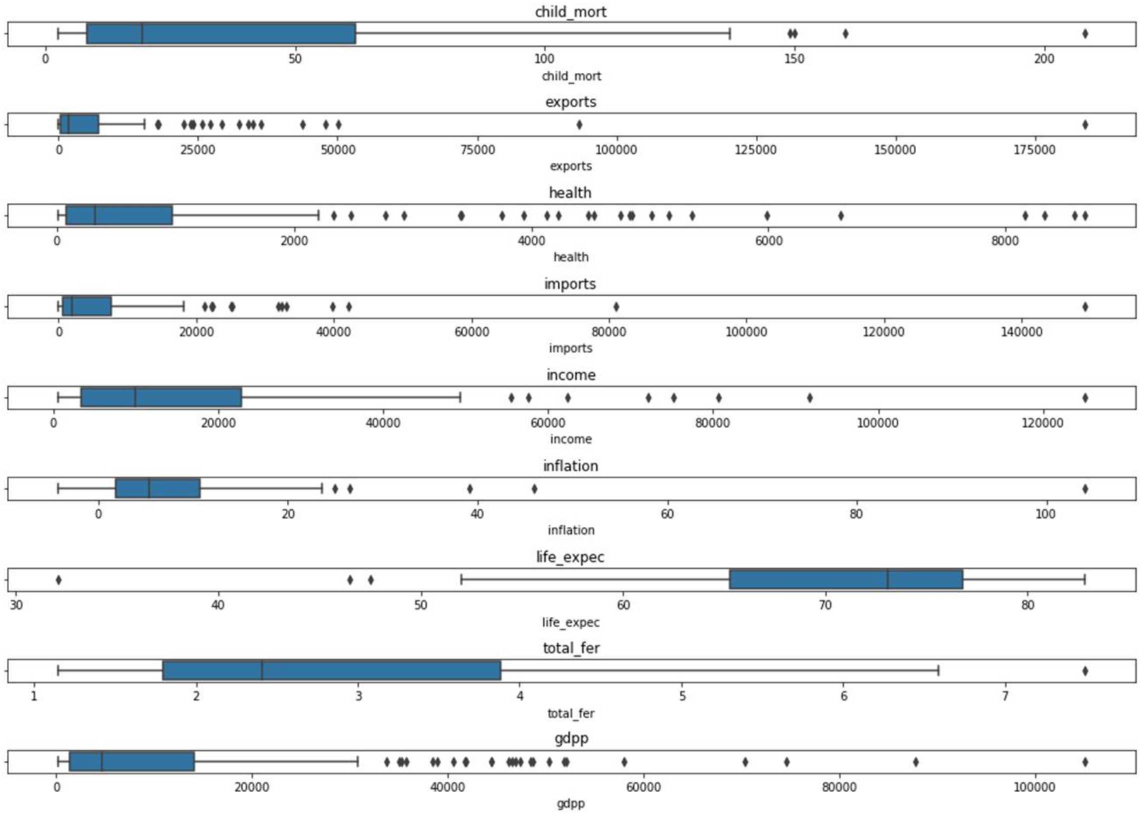

Reading & Understandin g Data
Data Preparation and EDA
Principal Component Analysis
Clustering (Kmeans and Hierarchical)
Cluster Analysis

# Overall Approach

- Reading & Understanding Data
- Cleaning the Data
  - Checking Missing Values
- Data Preparation & EDA
  - Handling Percentage Formats
  - Outlier Identification & Treatment
  - Correlation of Variables
- Principal Component Analysis
  - Scaling
  - Applying PCA and Plotting Principal Components
  - Scree Plot
  - Incremental PCA
  - Outlier Analysis on Principal Components
- Clustering
  - Hopkin Statistic
  - SSD and Silhouette Score
  - KMeans Clustering
  - Hierarchical Clustering
- Analysis of the Clusters
  - Analyzing Features in Clusters
  - Feature Means in Cluster
  - List of Countries in Dire-Need of Aid

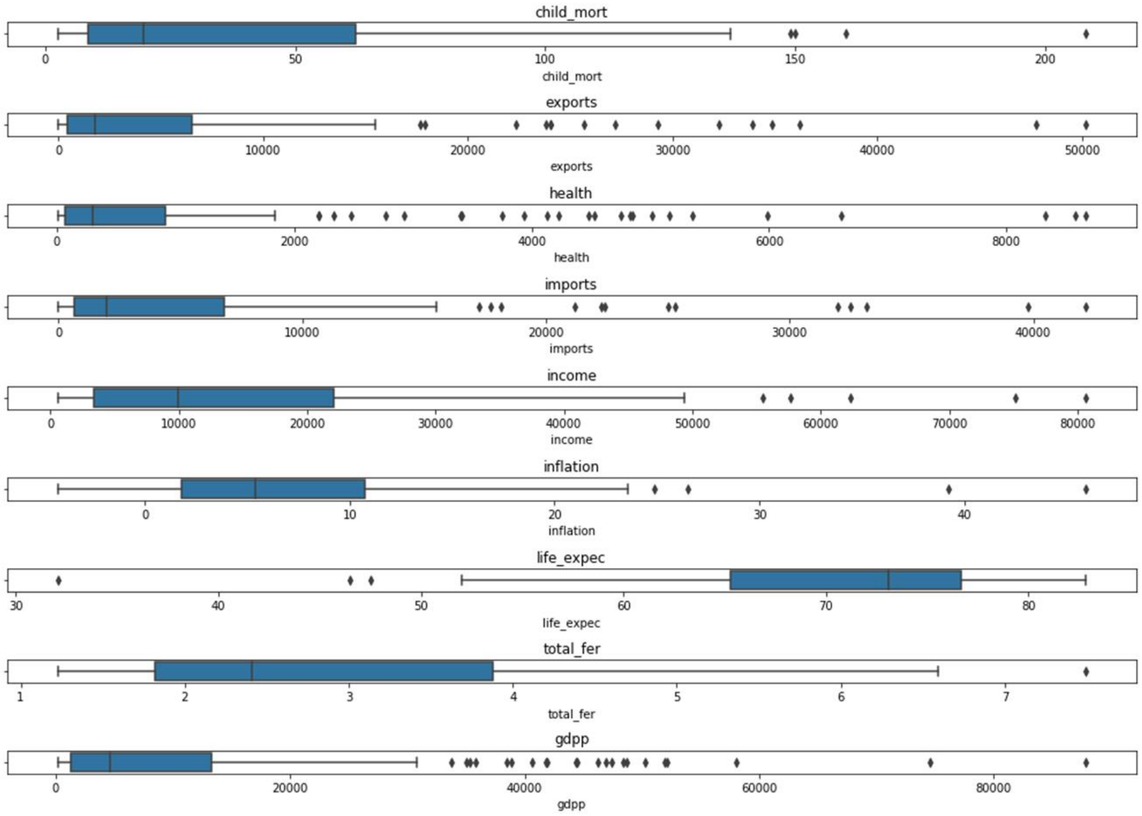
- Reading & Understanding Data
- Cleaning the Data
- Data Preparation & EDA

# Outlier Analysis & Treatment



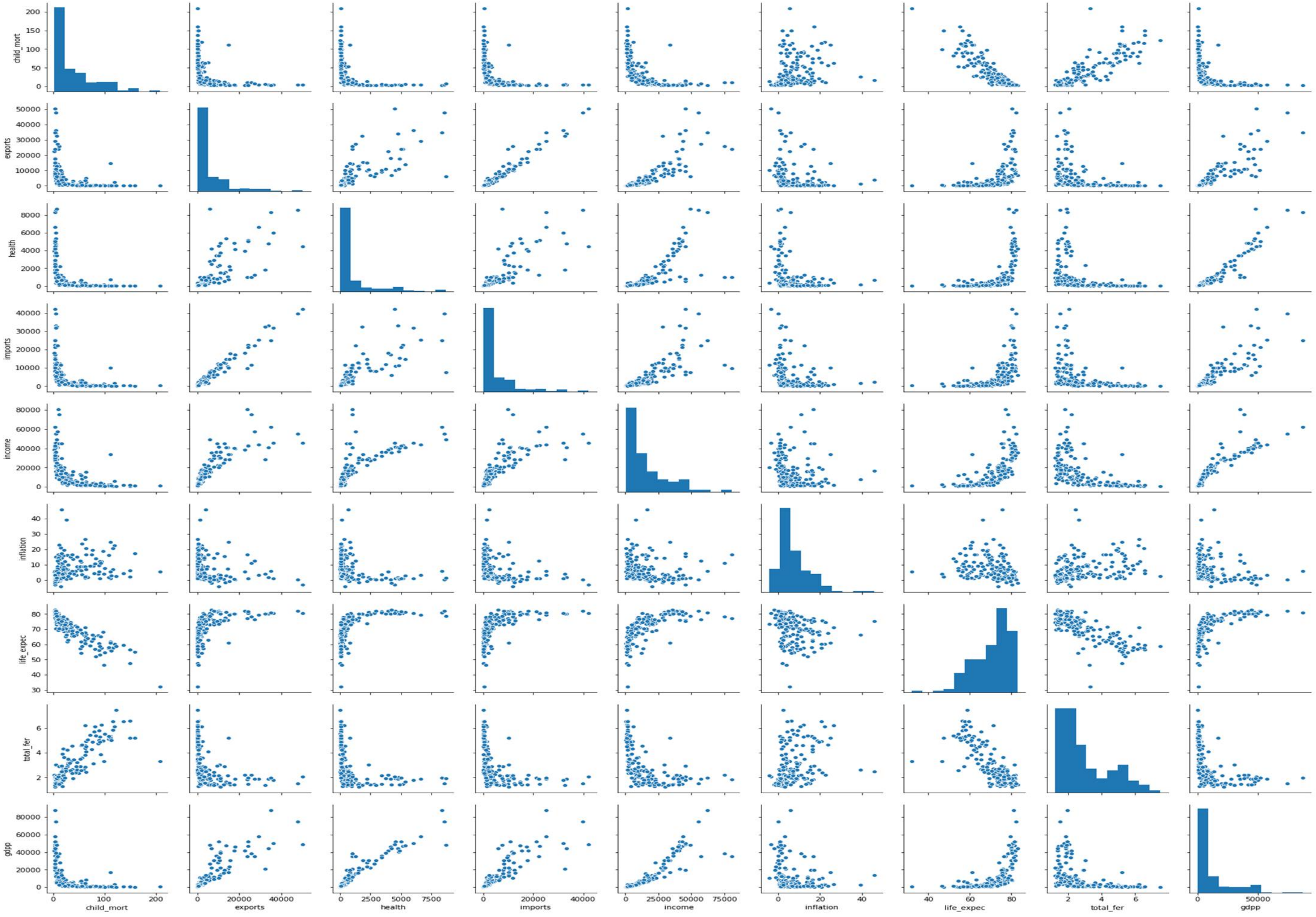
- Values beyond 75th and 85th percentile but the data seems all valid and logical.
- Most of the **outliers seems to be for developed countries** having very high gdpp and income and hence treating the outliers will not remove the countries that are under-developed and require aid and will not impact the objective.
- Treating the outliers by removing data below 5th percentile and beyond 95th percentile

# Outlier Treatment

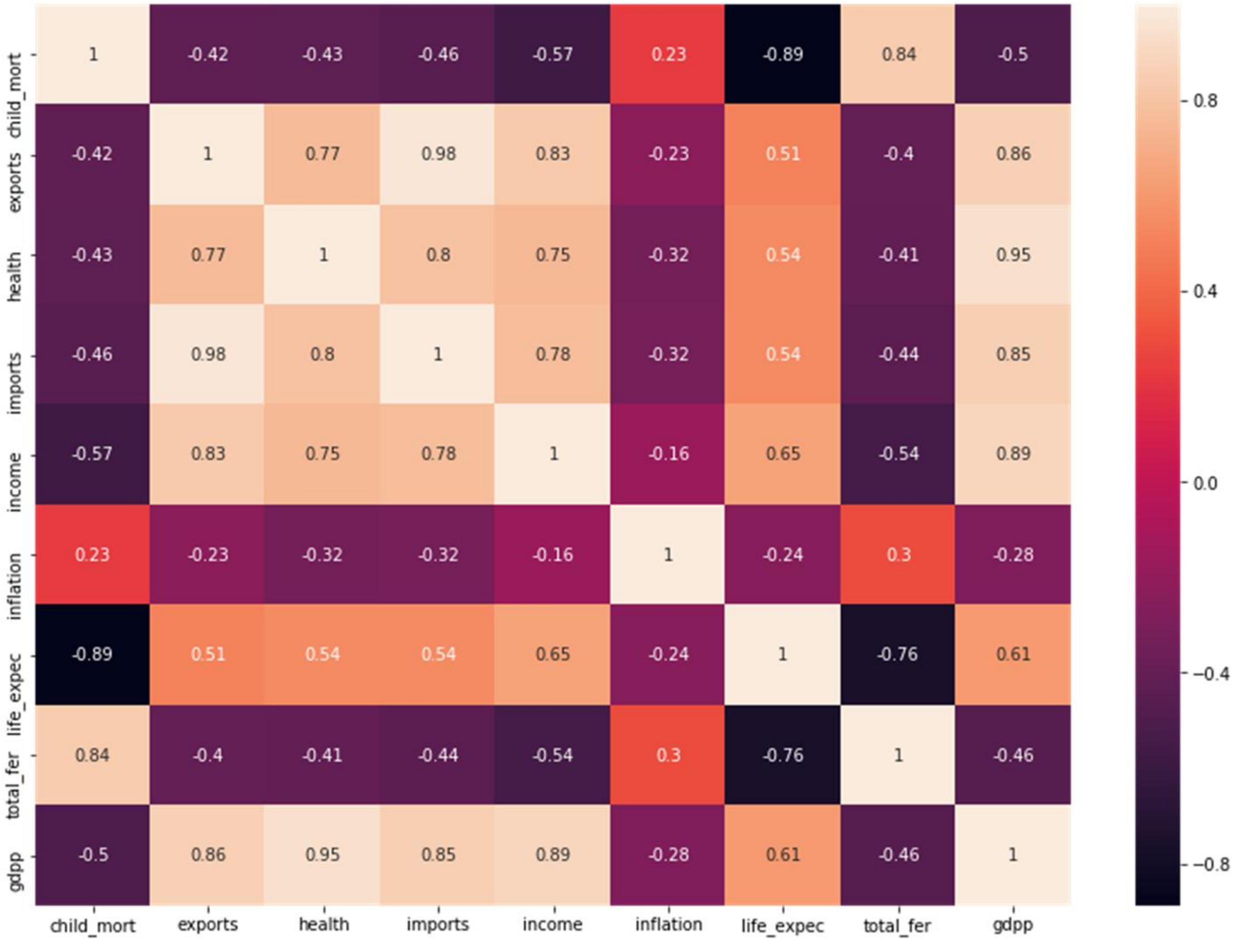


- 4 rows removed from the dataset as a result of outlier treatment
- From the box plots it is visible that the outlier for highly developed countries was removed. Based on our business problem that does not impact any under-developed country being removed from the dataset

# Correlation of Variables (Pair Plot)



# Correlation of Variables (Heat Map)



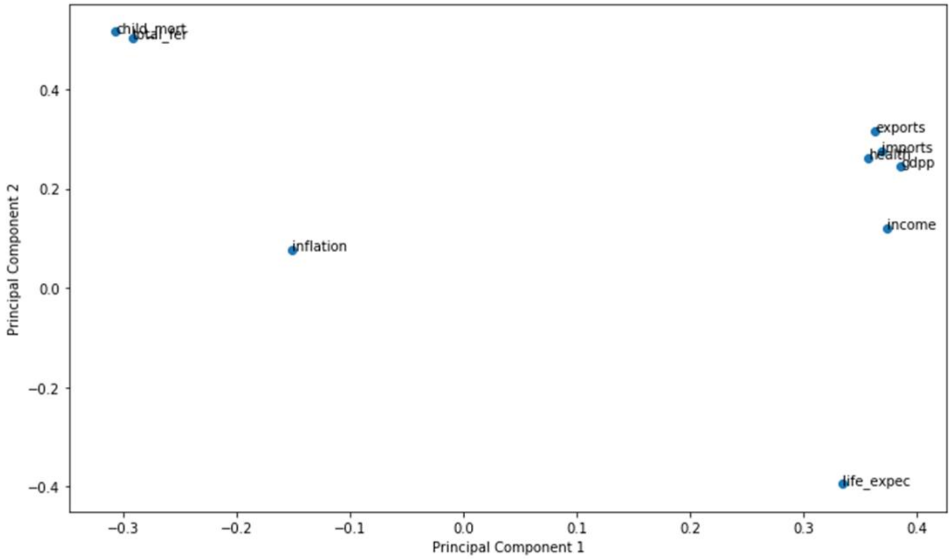
| | child_mort | exports | health | imports | income | inflation | life_expec | total_fer | gdpp |
|---|---|---|---|---|---|---|---|---|---|
| child_mort | 1 | -0.42 | -0.43 | -0.46 | -0.57 | 0.23 | -0.89 | 0.84 | -0.5 |
| exports | -0.42 | 1 | 0.77 | 0.98 | 0.83 | -0.23 | 0.51 | -0.4 | 0.86 |
| health | -0.43 | 0.77 | 1 | 0.8 | 0.75 | -0.32 | 0.54 | -0.41 | 0.95 |
| imports | -0.46 | 0.98 | 0.8 | 1 | 0.78 | -0.32 | 0.54 | -0.44 | 0.85 |
| income | -0.57 | 0.83 | 0.75 | 0.78 | 1 | -0.16 | 0.65 | -0.54 | 0.89 |
| inflation | 0.23 | -0.23 | -0.32 | -0.32 | -0.16 | 1 | -0.24 | 0.3 | -0.28 |
| life_expec | -0.89 | 0.51 | 0.54 | 0.54 | 0.65 | -0.24 | 1 | -0.76 | 0.61 |
| total_fer | 0.84 | -0.4 | -0.41 | -0.44 | -0.54 | 0.3 | -0.76 | 1 | -0.46 |
| gdpp | -0.5 | 0.86 | 0.95 | 0.85 | 0.89 | -0.28 | 0.61 | -0.46 | 1 |

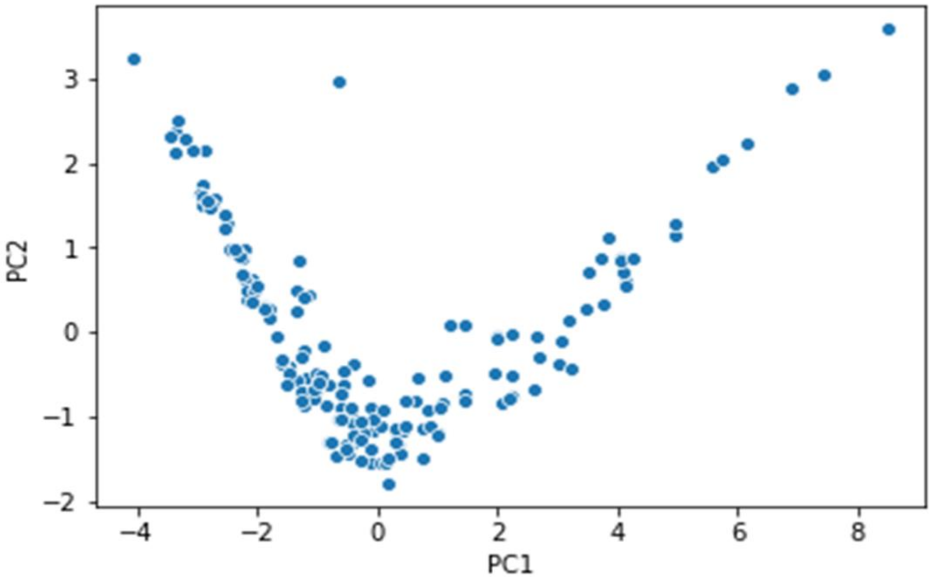
- (exports and imports), (health and gdp), (income and gdpp), (child_mort and life_expec), (exports and gdpp), (imports and gdpp) and (child_mort and total_fer) have the highest correlation in the dataset

- Principal Component Analysis

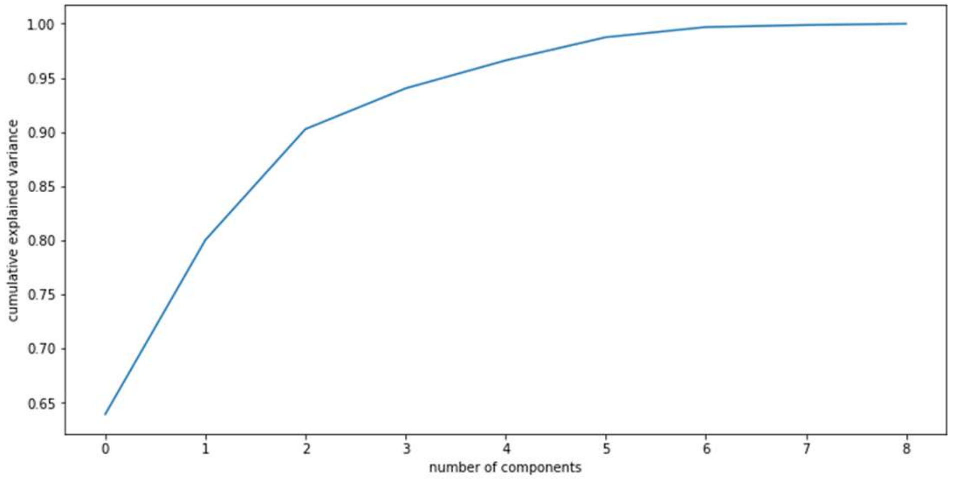
# Scaling, Variance & Visualization

- Scaling - Scaled the dataset
- Variance - Computed Explained Variance Ratio to determine the how many Principal Components best explain the dataset
- Visualizing 2 Principal Components
  - The first component is where the income, gdpp, life_expec, imports, exports and health is heavy. These 6 components have the highest loading.
  - The second component is where the child_mort, total_fer are heavy. These 2 components have the highest loading.





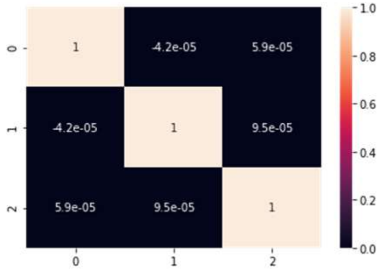
# Scree Plot, Correlation & Outlier Analysis



| Cumulative Explained Variance Ratio |
|---|
| 0.639689 (1 PC) |
| 0.800416 (2 PC) |
| 0.902788 (3 PC) |
| 0.940345 (4 PC) |
| 0.966142 (5 PC) |
| 0.987509 (6 PC) |
| 0.996912 (7 PC) |
| 0.998783 (8 PC) |
| 1.000000 (9 PC) |

- Around 80% variance is explained by 2 components
- Around 90% variance is explained by 3 components
- Around 94% variance is explained by 4 components

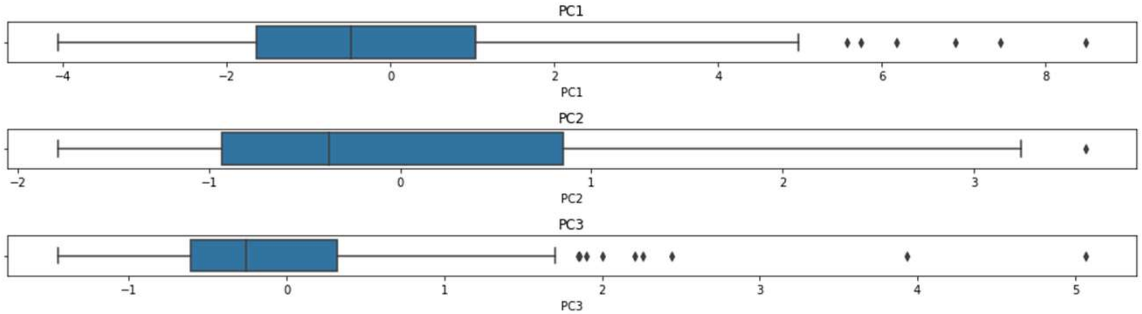
**Choosing 3 principal components for our model**



| | 0 | 1 | 2 |
|---|---|---|---|
| 0 | 1 | -4.2e-05 | 5.9e-05 |
| 1 | -4.2e-05 | 1 | 9.5e-05 |
| 2 | 5.9e-05 | 9.5e-05 | 1 |

### Correlation Matrix

- There is no correlation between these three components. They are orthogonal.
- We have effectively removed multicollinearity



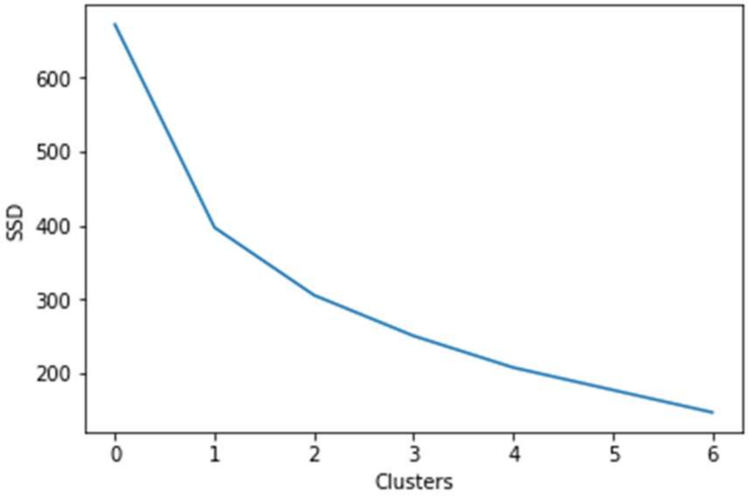
### Outlier Analysis on Principal Components

- Principal components do have outliers but the data does not seem to outlie significantly.
- Not removing any outliers on principal components.
- They were earlier treated in the main dataset.

- Clustering

# Clustering

- Hopman Statistics of 87% suggesting the dataset has a good tendency to form clusters
- Finding Optimal Number of Clusters



| Clusters | Average Silhouette |
|---|---|
| 2 | 0.50978 |
| 3 | 0.45211 |
| 4 | 0.4046 |
| 5 | 0.4200 |
| 6 | 0.3784 |
| 7 | 0.3887 |
| 8 | 0.3920 |

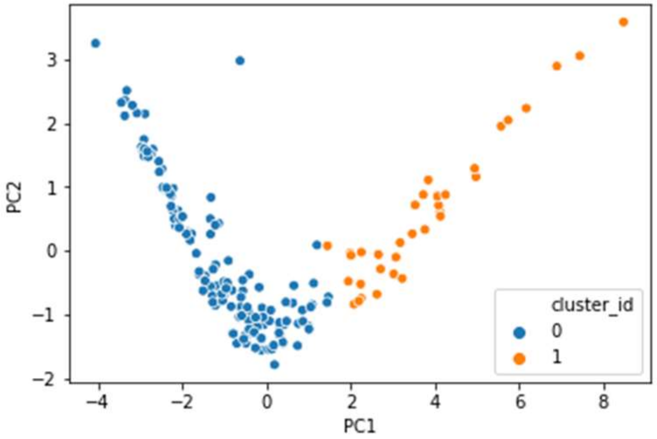
## SSD (Elbow Curve Inference)

- Based on the elbow bend 2 or 3 clusters seem to suffice our need

## Average silhouette Interpretation

- A Silhouette score of 0.51 is coming for 2 clusters and 0.45 for 3 clusters

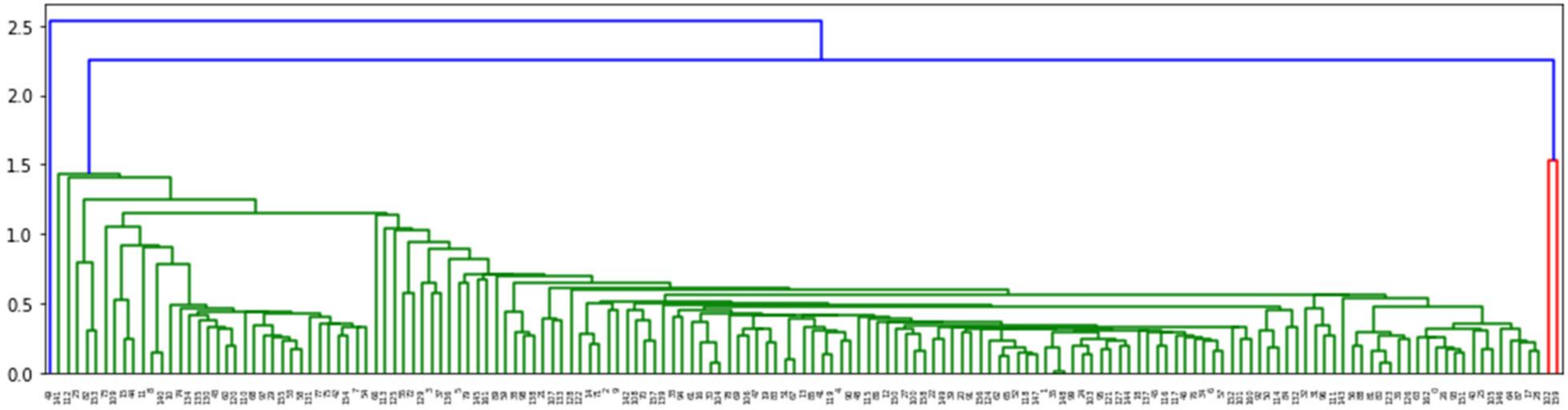
**Choosing 2 clusters for our model based on both SSD and Silhouette Score**



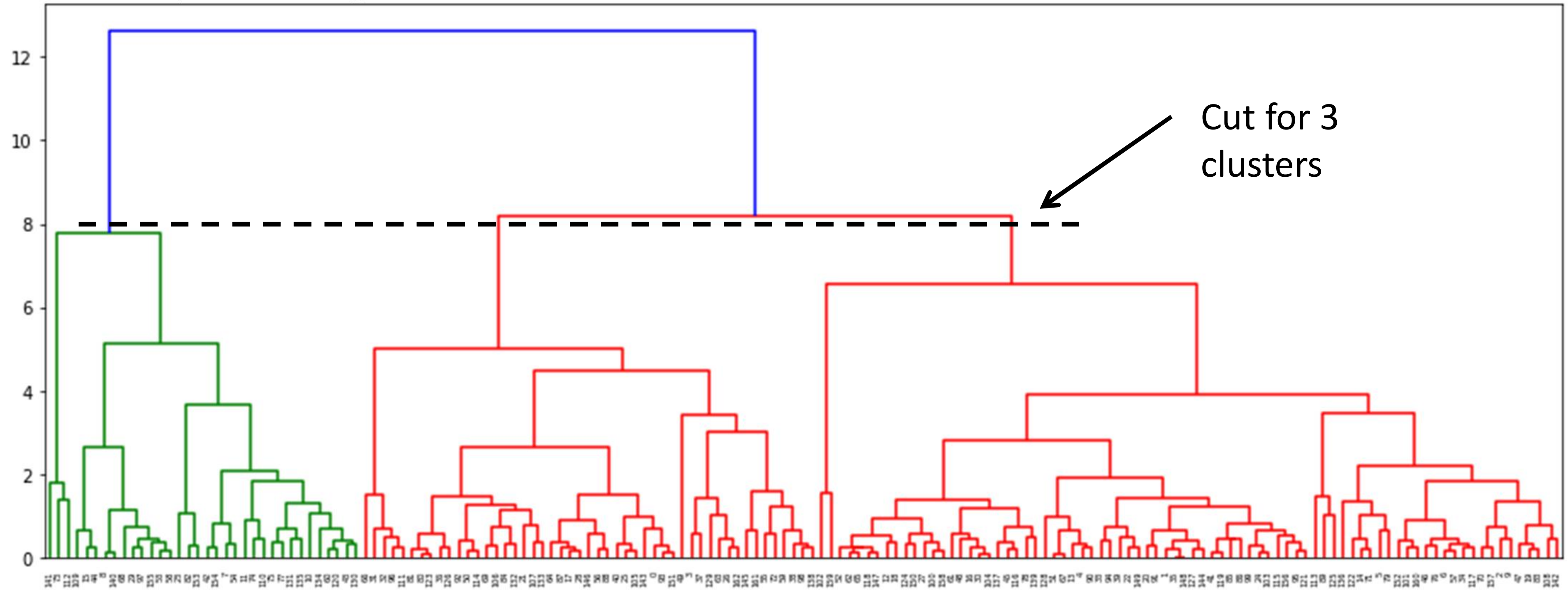
*Visualizing 2 Principal Components for 2 Clusters*

# Hierarchical Clustering

- Single Linkage Plot



- Complete Linkage Plot



Cut for 3 clusters

# Hierarchical Clustering

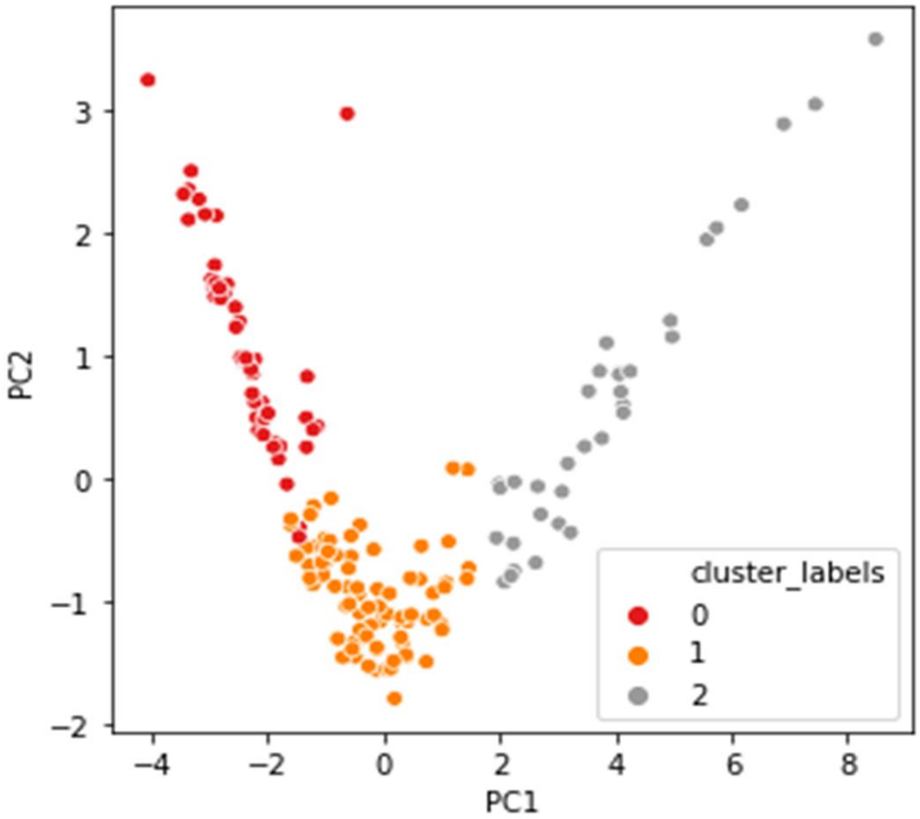
- Visually, Complete linkage dendrogram, shows 3 to 4 possible clusters
- From a four cluster business perspective we could possibly derive :
  - Developed (High)
  - Developing (Upper-Middle)
  - Developing (Lower-Middle)
  - Under-Developed (Low)

  One of the Clusters will only have three data points, hence not useful.
- From a three cluster business perspective it could possibly derive to:
  - **Developed Countries**
  - **Developing Countries**
  - **Under-Developed Countries**
- Objective is to determine under-developed countries requiring direct aid there is not much benefit in dividing the cluster of developing countries into two separate clusters.

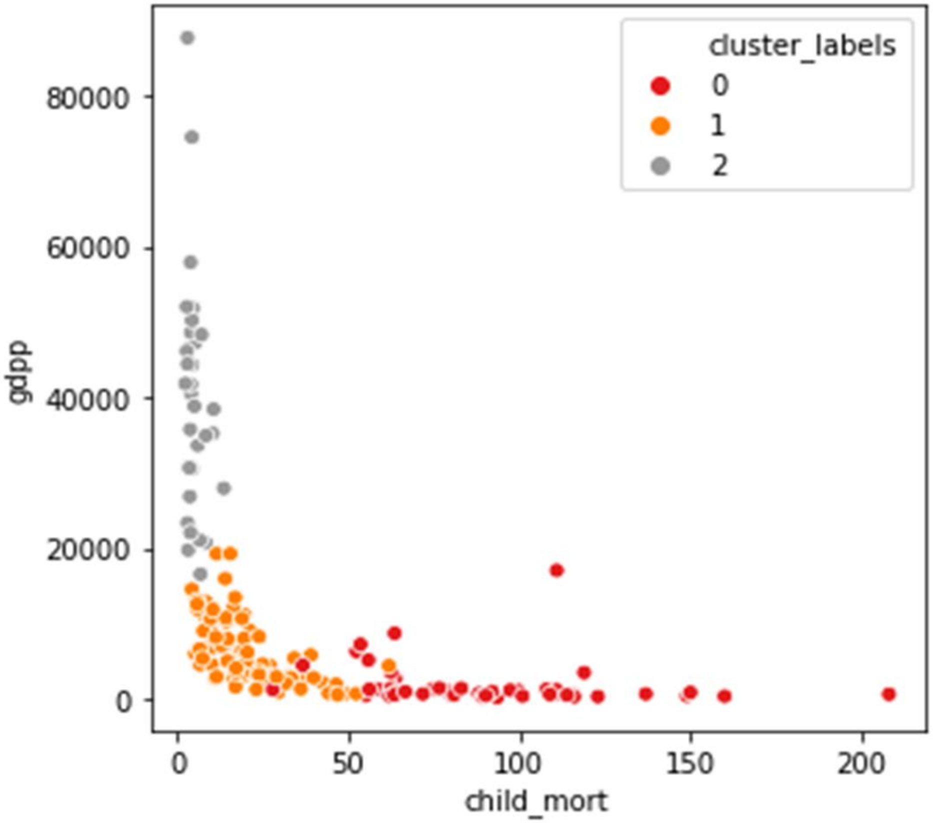
**Hence, selecting 3 clusters**

# ❖ Cluster Analysis

# Cluster Analysis and Visualizations

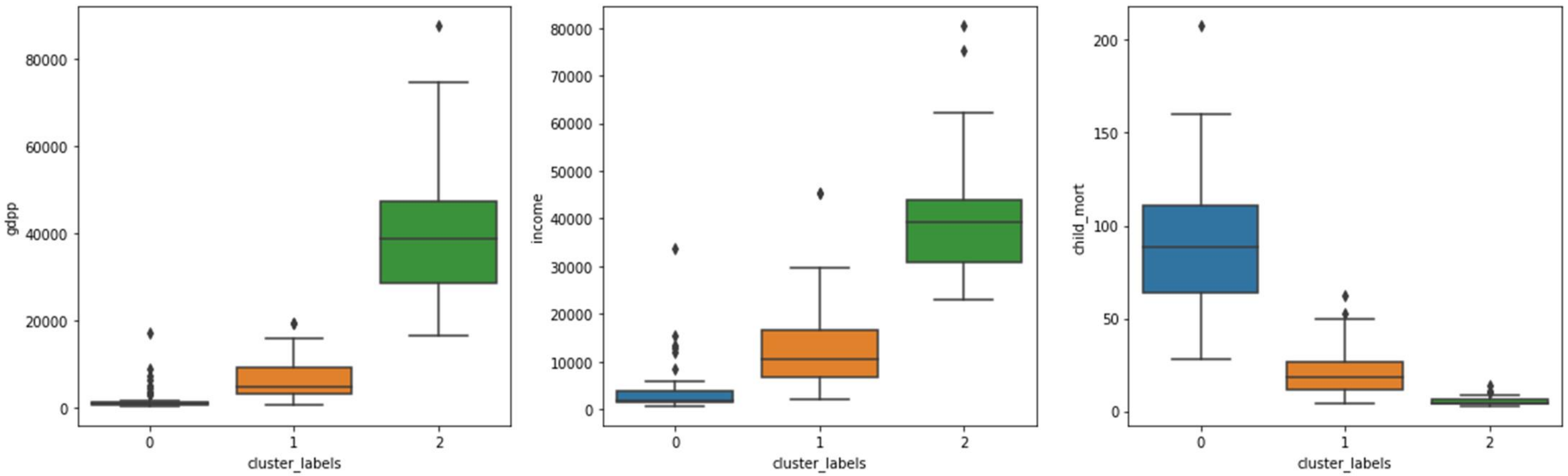


- Principal Components (**PC1 and PC2**) Cluster Visualization
- Clear separation for Developed, Developing and Under-Developed Countries



- **gdpp and child_mort** features - Cluster Visualization
- Clear separation for Developed, Developing and Under-Developed Countries

# Analyzing Features in Clusters



- Higher gdpp and income in Cluster 2 suggests that **Cluster 2 is for developed countries**
- Moderate gdpp and income in Cluster 1 suggests that **Cluster 1 is for developing countries**
- Higher child_mort and low gdpp and income in Cluster 0 suggests that **Cluster 0 is for under-developed countries**

# Feature Means in Clusters

| Cluster | child_mort | exports | health | imports | income | inflation | life_expec | total_fer | gdpp |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 89.60 | 855.34 | 111.58 | 809.36 | 3878.63 | 9.87 | 59.51 | 4.87 | 1870.41 |
| 1 | 20.96 | 2855.06 | 401.96 | 2968.42 | 12510.75 | 7.70 | 73.10 | 2.28 | 6415.42 |
| 2 | 5.27 | 19046.77 | 3748.69 | 16935.36 | 40538.24 | 2.51 | 79.85 | 1.77 | 39291.18 |

- child_mort, inflation and total_fer is highly loaded in Cluster 0.
  - Signifying this cluster is of under-developed countries.
- All variables are in the mid-range in Cluster 1.
  - Signifying this cluster is for developing countries.
- gdpp, income, exports, imports, health and life_expec is highly loaded in Cluster 2.
  - Signifying this cluster is of developed countries.

**Inference: Cluster 0 is in dire-need of aid**

# Conclusion

Filtering Countries in Cluster 0

- Countries above means of child_mort, inflation and total_fer
- And countries below means of gdpp, income, health, exports, imports and life_expec

## List of Countries in Dire-Need of Aid

- Burundi
- Congo, Dem. Rep.
- Sierra Leone
- Malawi
- Guinea