Machine Learning1

Cluster-PCA Assignment)

IIITB Data Science Course 2

By: Hari Thapliyal

Problem Statement

The dataset given has financial, health, population growth, mortality related data of 167 countries of the world. HELP international need to allocate $10 million fund to the countries, who deserves the most. The challenges before us is to identify which 5 countries deserve this fund the most.

We should cluster these countries based on the given data. After we are done with clustering we need to select 5 countries from the created clusters.

Second problems which I tried to solve is if some new country is added in this dataset or values of given parameters of the existing country change then predict in which cluster that particular country should be placed.

Analysis Approach

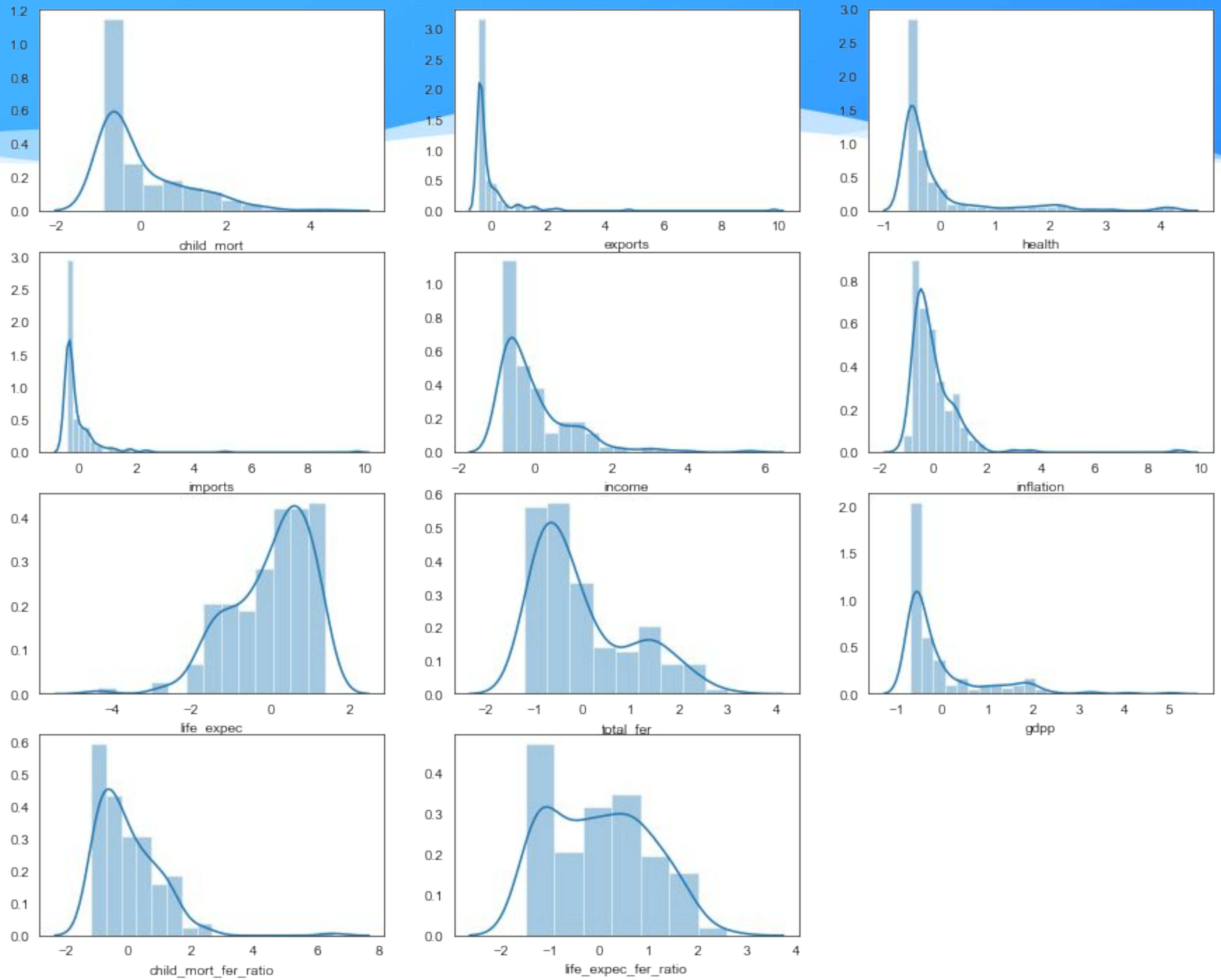
1. Creating meaningful fields from % fields
2. Creating 2 more useful fields - Child_mortality_Fertality_Ratio, Life_expectency_Fertality_Ratio.
3. **Scale the data** using standard scaler
4. Perform PCA and get PCA features
5. Perform Hopkins Analysis- **Data is good** for Clustering?
6. Scree plot to Identify **how many PCA** will be sufficient to accommodate 90%+ variance
7. Visualize contribution of each original feature in important PCAs
8. Create a Dendogram
9. Perform silhoutte analysis, plot Elbow curve to determine **how many clusters** will be required
10. **Decide the number of cluster** required and assign those to PCA and original data
11. Determine the mean of gdpp, mortality, income for each cluster
12. Plot countries using PCA and important variables
13. Identify 5 countries which need the funding

Results of Principal Component Analysis and Clustering

Identified 4 PCA which can represent 91% of the variation in the dataset

Visualizations of the most Important Results

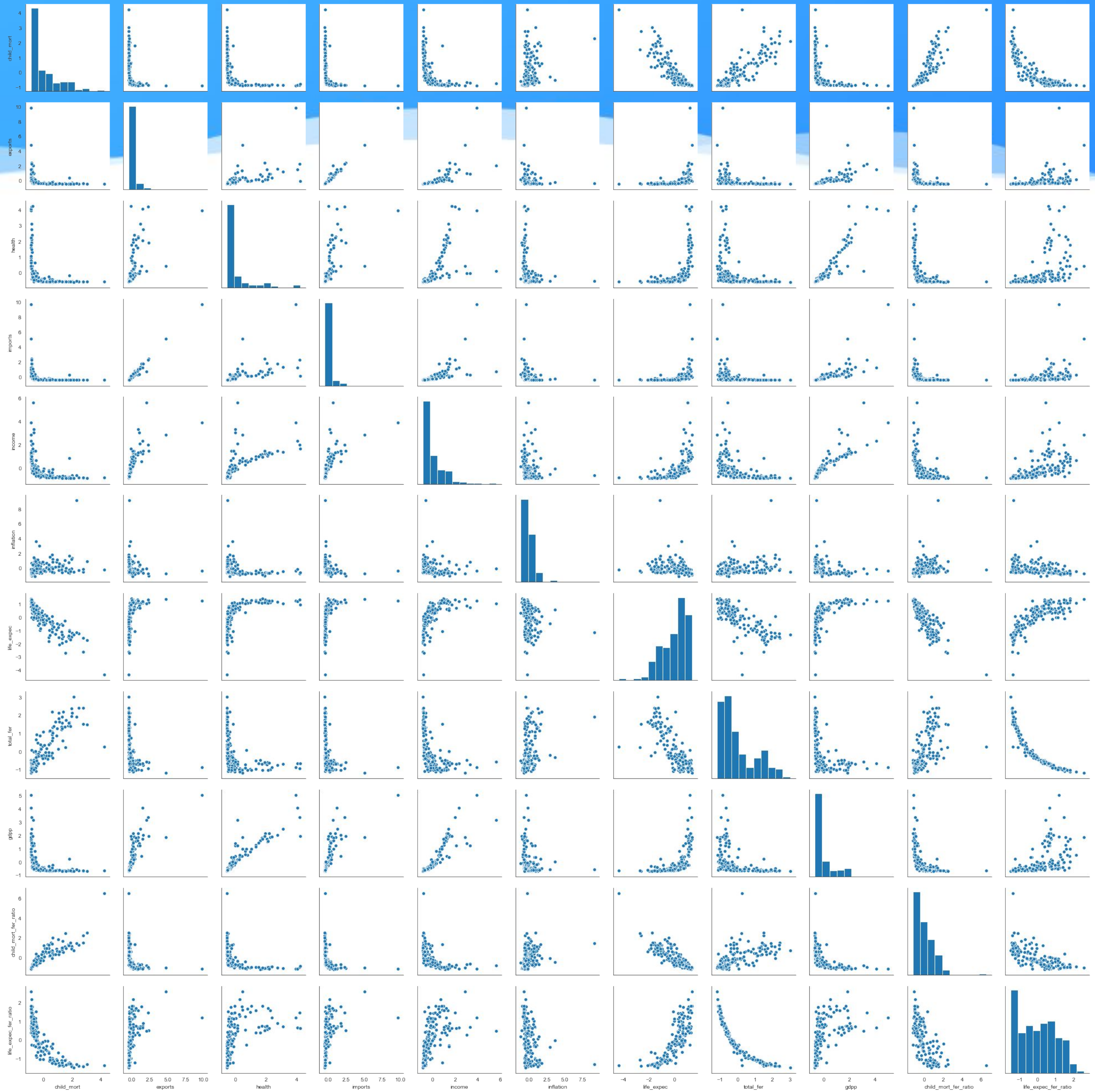
Data Distribution



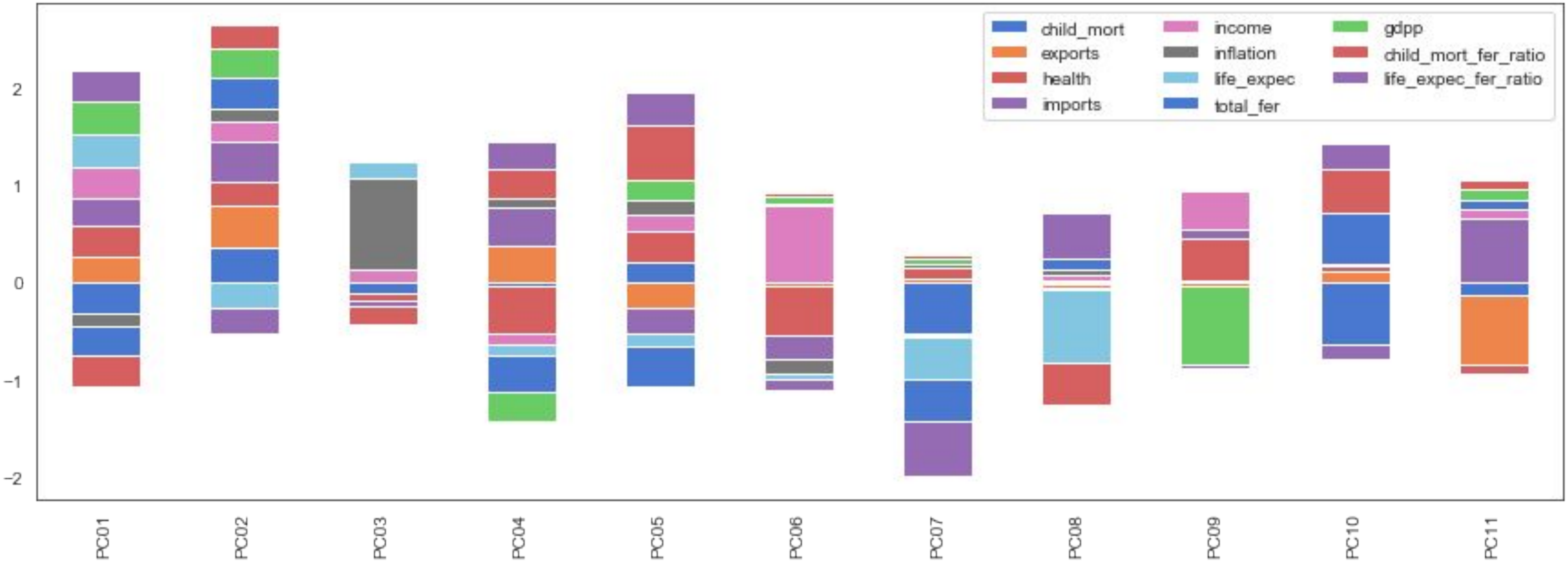
Correlation and Shape of Relation

From Top-bottom, Left-Right Variables name in this correlation graph

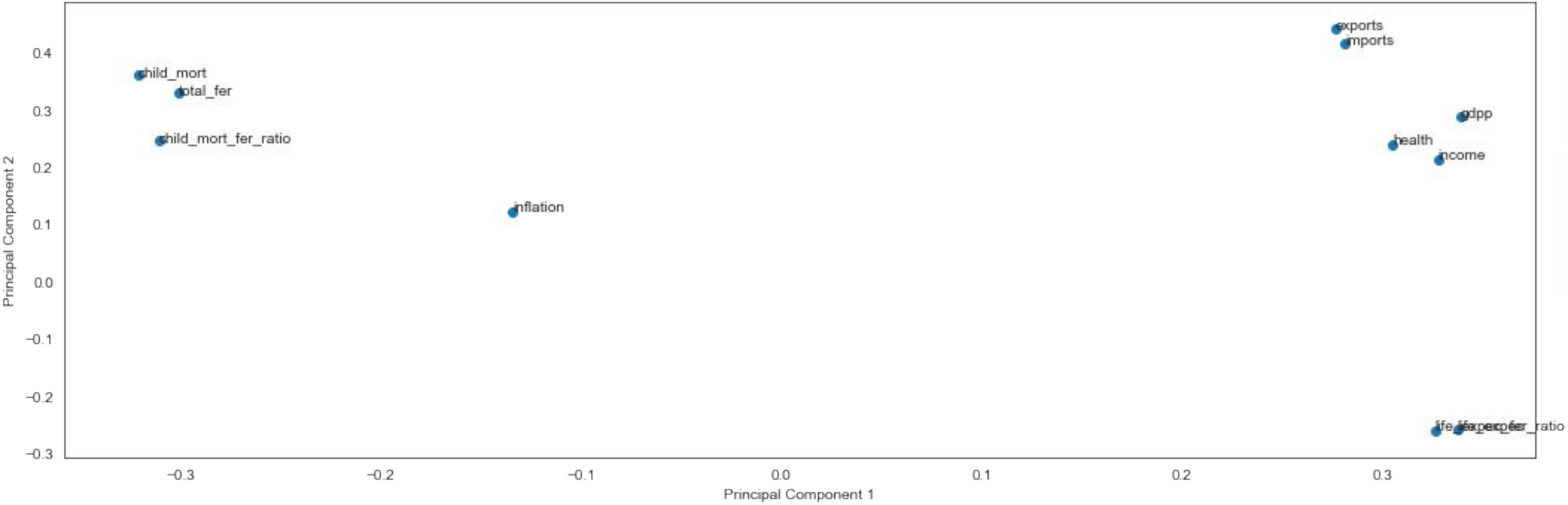
1. child_mort
2. exports
3. health
4. imports
5. income
6. inflation
7. life_expec
8. total_fer
9. gdpp
10. child_mort_fer_ratio
11. life_expec_fer_ratio



Contribution of Each Feature in PCA



PC1 & PC2 Contribution of Original Features

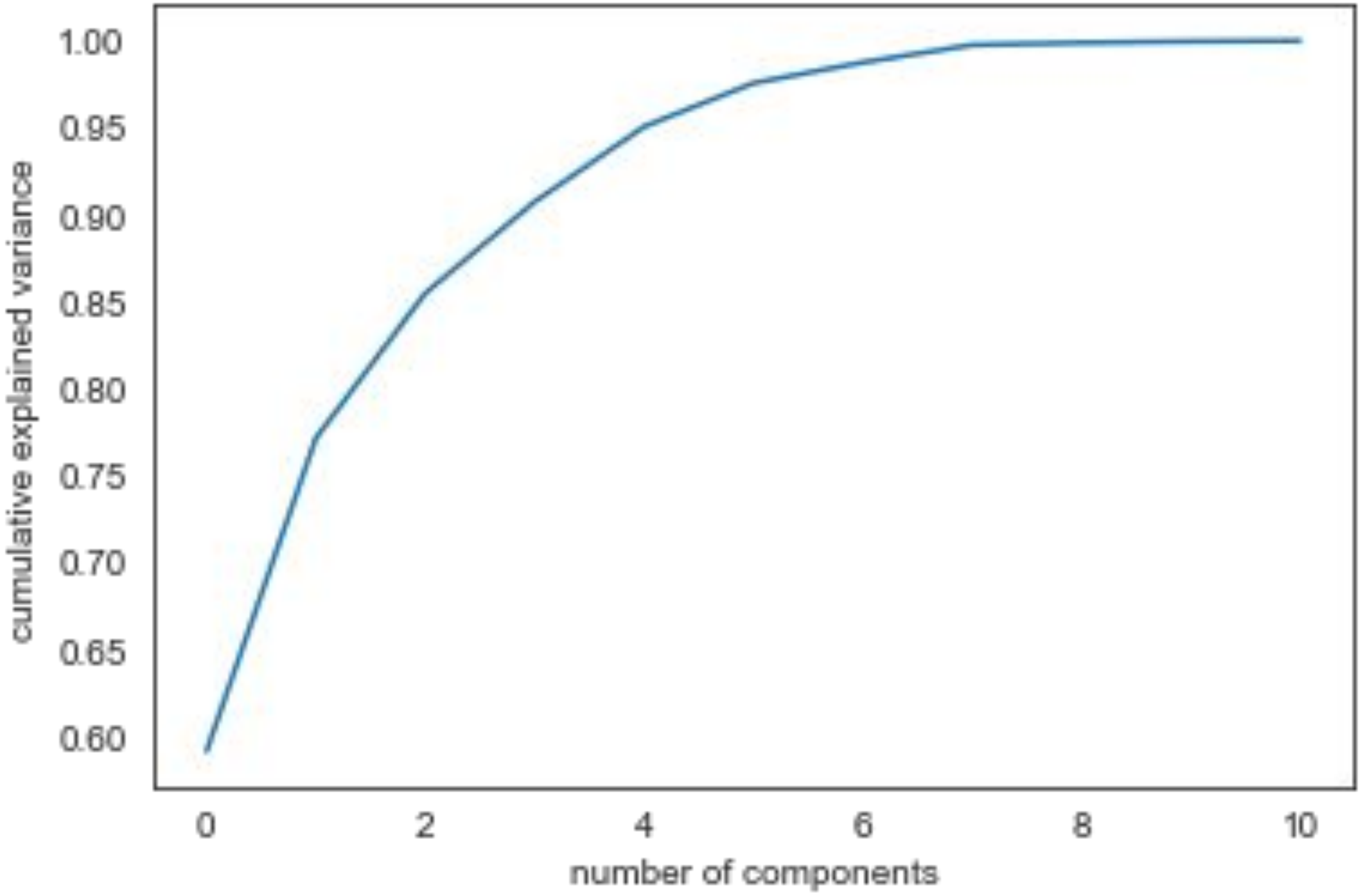


Screeplot

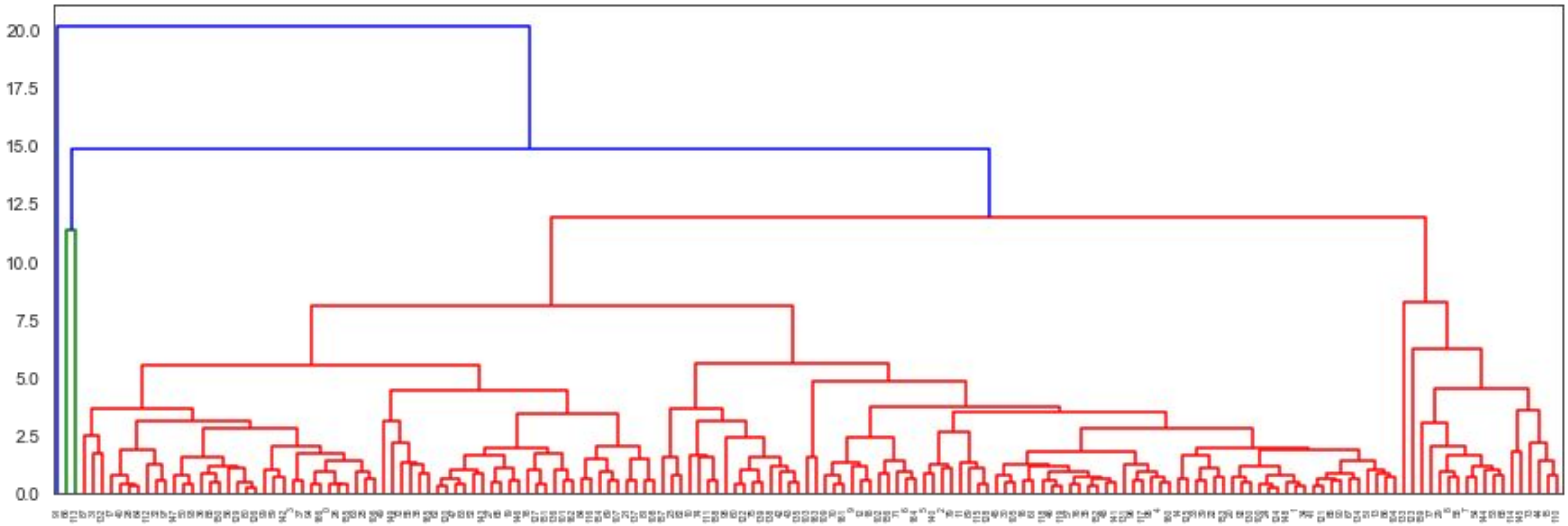
Screeplot –

Plotting the cumulative variance against the original number of components

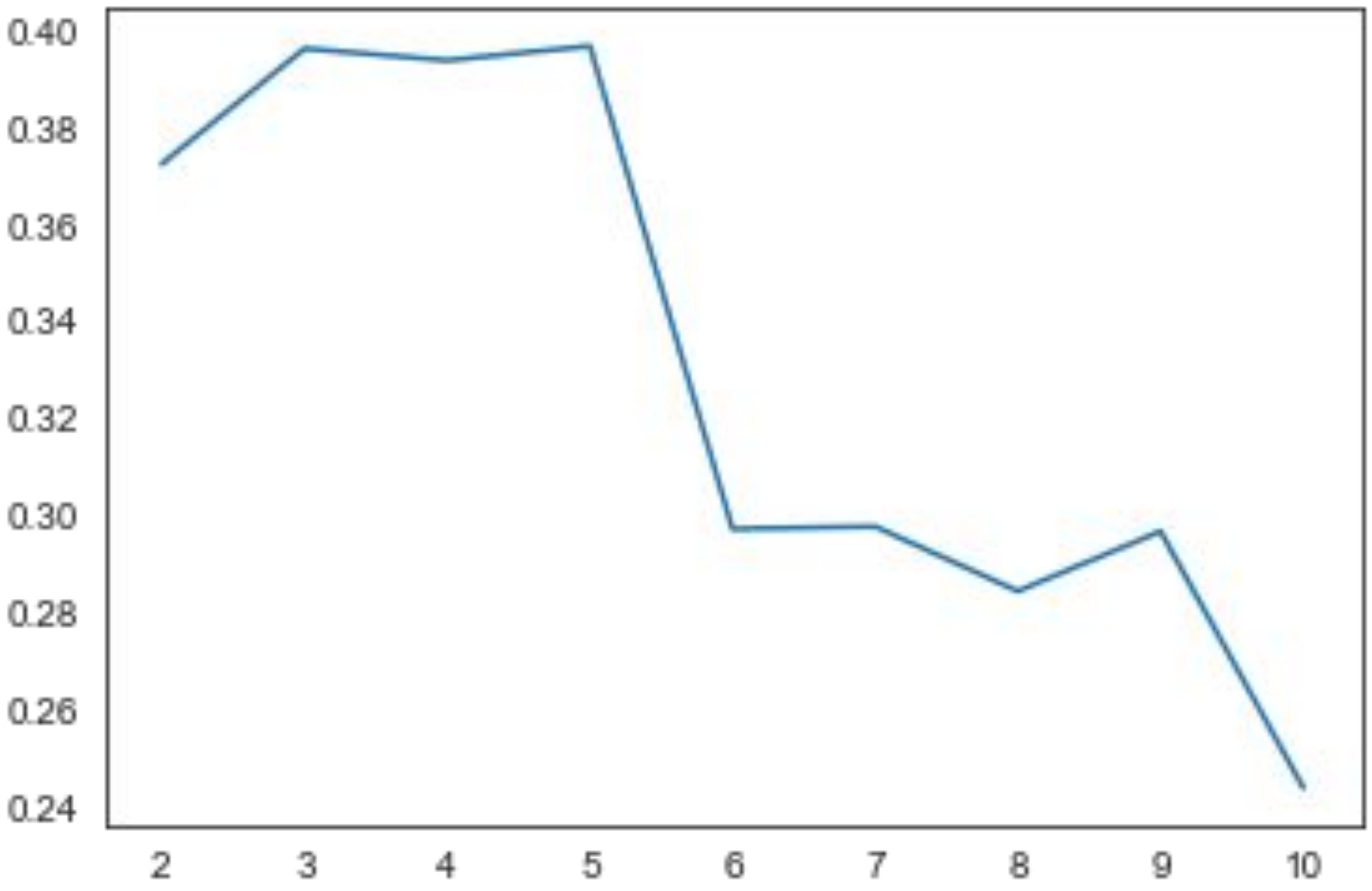
Shows 4 PCA can explain 91 of variables



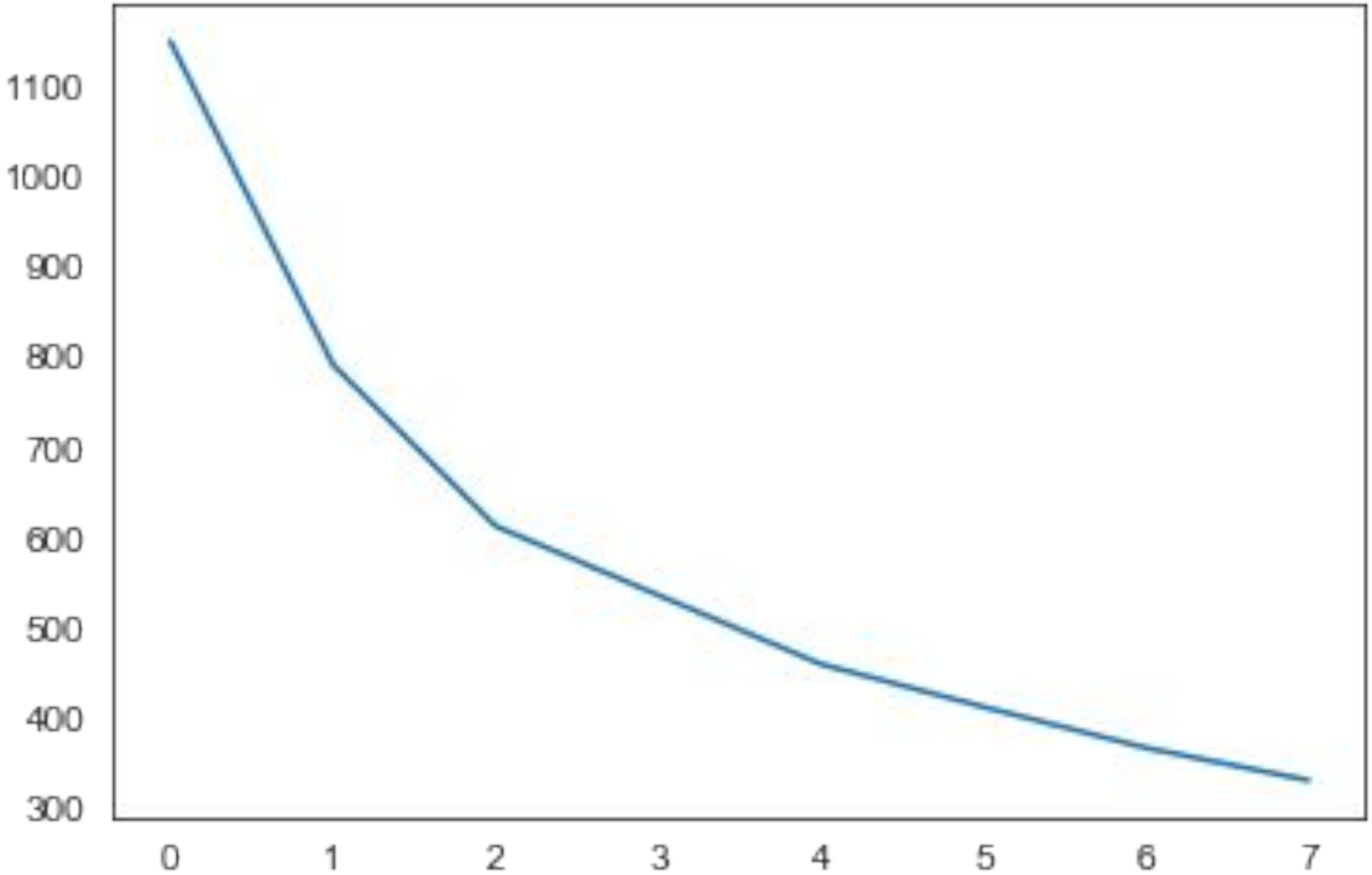
Dendogram Using PCA (Complete)



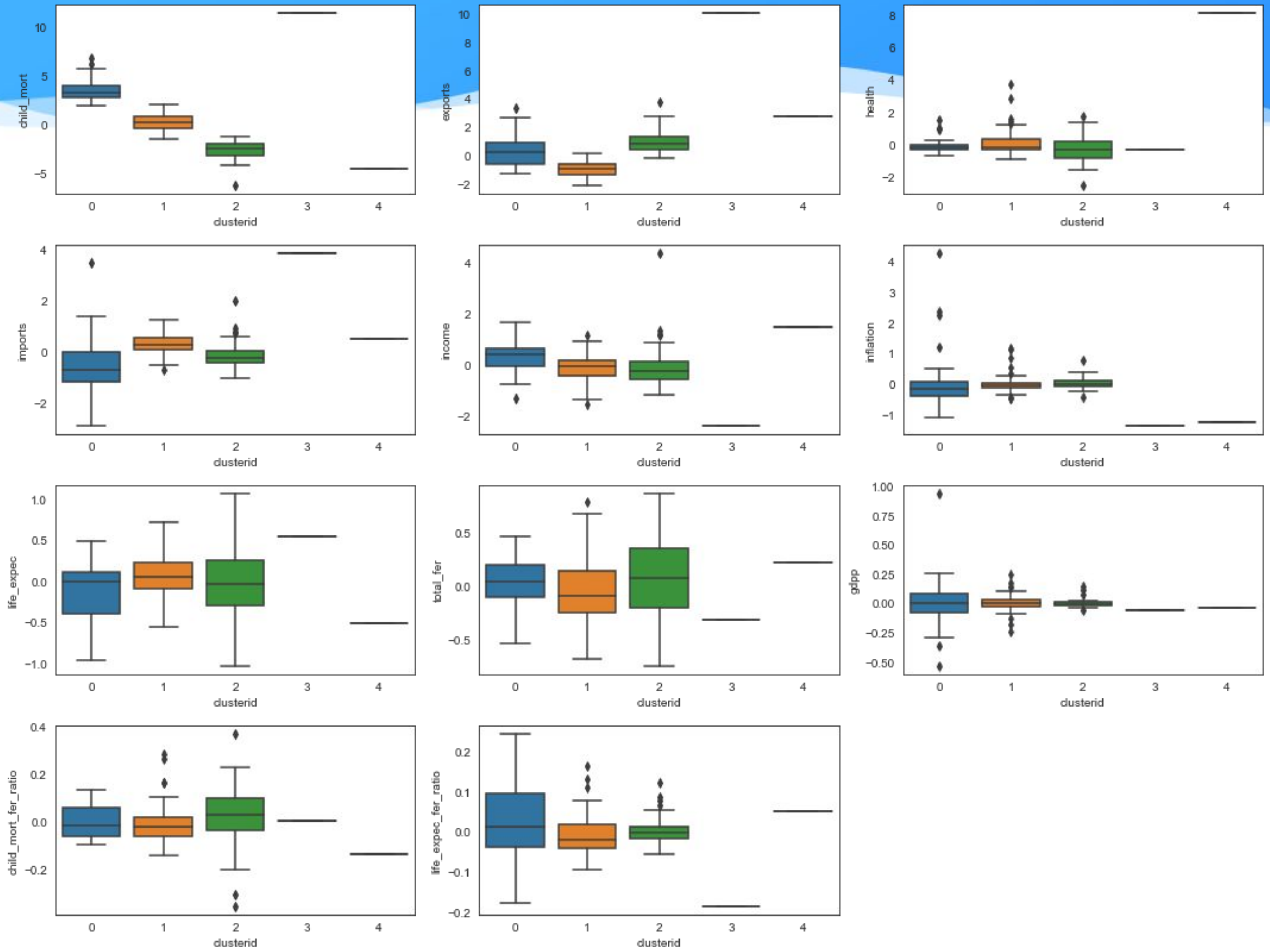
Silhouette Score



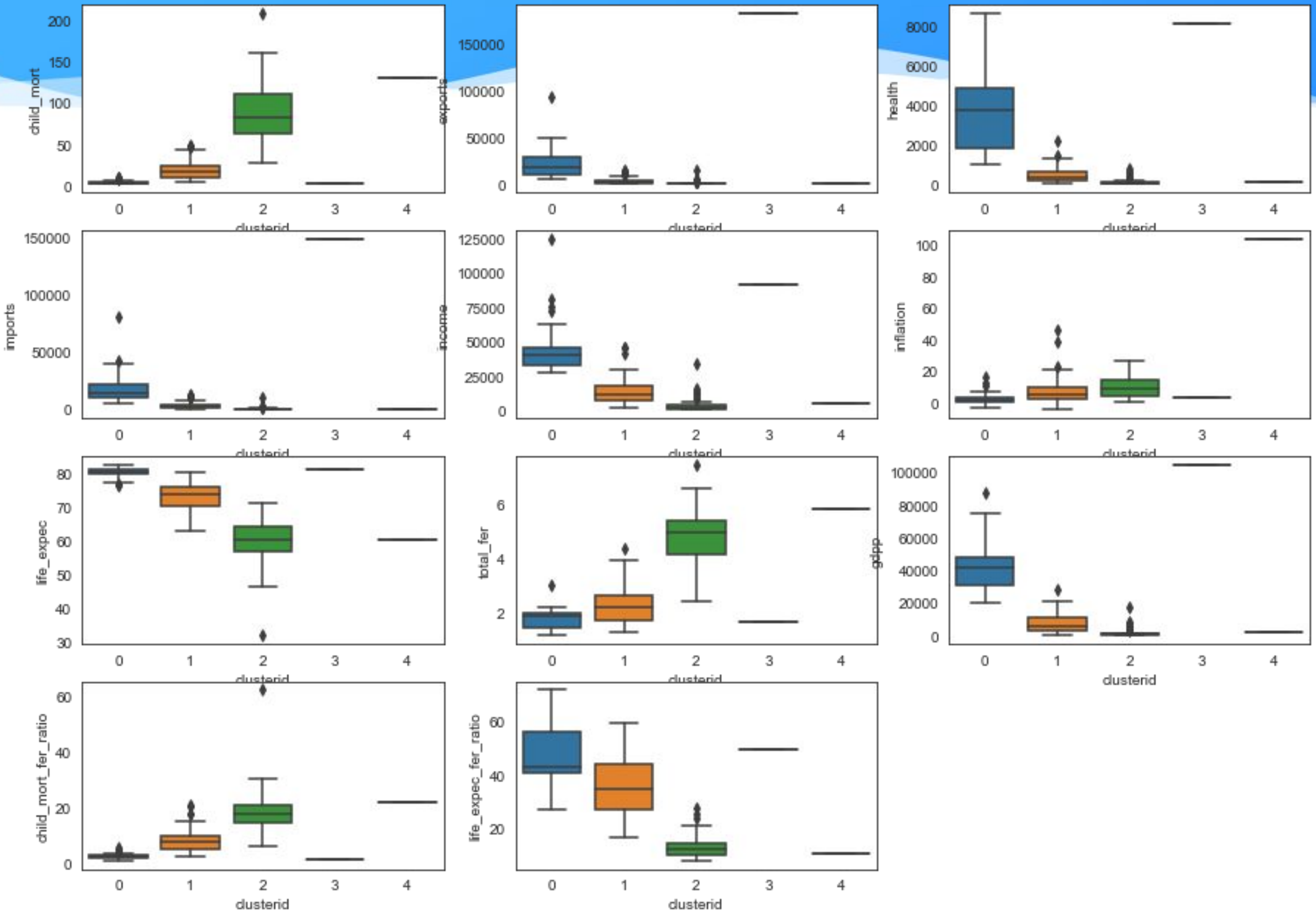
Elbow Curve



Distribution using PCA 5 Clusters



Distribution using Actual Data
7 Clusters



Countries in Clusters

- Clusterid : 0

  Australia, Austria, Belgium, Brunei, Canada, Cyprus, Czech Republic, Denmark, Finland, France, Germany, Greece, Iceland, Ireland, Israel, Italy, Japan, Kuwait, Malta, Netherlands, New Zealand, Norway, Portugal, Qatar, Singapore, Slovenia, South Korea, Spain, Sweden, Switzerland, United Arab Emirates, United Kingdom, United States

- Clusterid : 1

  Albania, Algeria, Antigua and Barbuda, Argentina, Armenia, Azerbaijan, Bahamas, Bahrain, Bangladesh, Barbados, Belarus, Belize, Bhutan, Bolivia, Bosnia and Herzegovina, Brazil, Bulgaria, Cambodia, Cape Verde, Chile, China, Colombia, Costa Rica, Croatia, Dominican Republic, Ecuador, Egypt, El Salvador, Estonia, Fiji, Georgia, Grenada, Guatemala, Guyana, Hungary, Indonesia, Iran, Jamaica, Jordan, Kazakhstan, Kyrgyz Republic, Latvia, Lebanon, Libya, Lithuania, Macedonia, FYR, Malaysia, Maldives, Mauritius, Micronesia, Fed. Sts., Moldova, Mongolia, Montenegro, Morocco, Nepal, Oman, Panama, Paraguay, Peru, Philippines, Poland, Romania, Russia, Samoa, Saudi Arabia, Serbia, Seychelles, Slovak Republic, Sri Lanka, St. Vincent and the Grenadines, Suriname, Thailand, Tonga, Tunisia, Turkey, Ukraine, Uruguay, Uzbekistan, Vanuatu, Venezuela, Vietnam

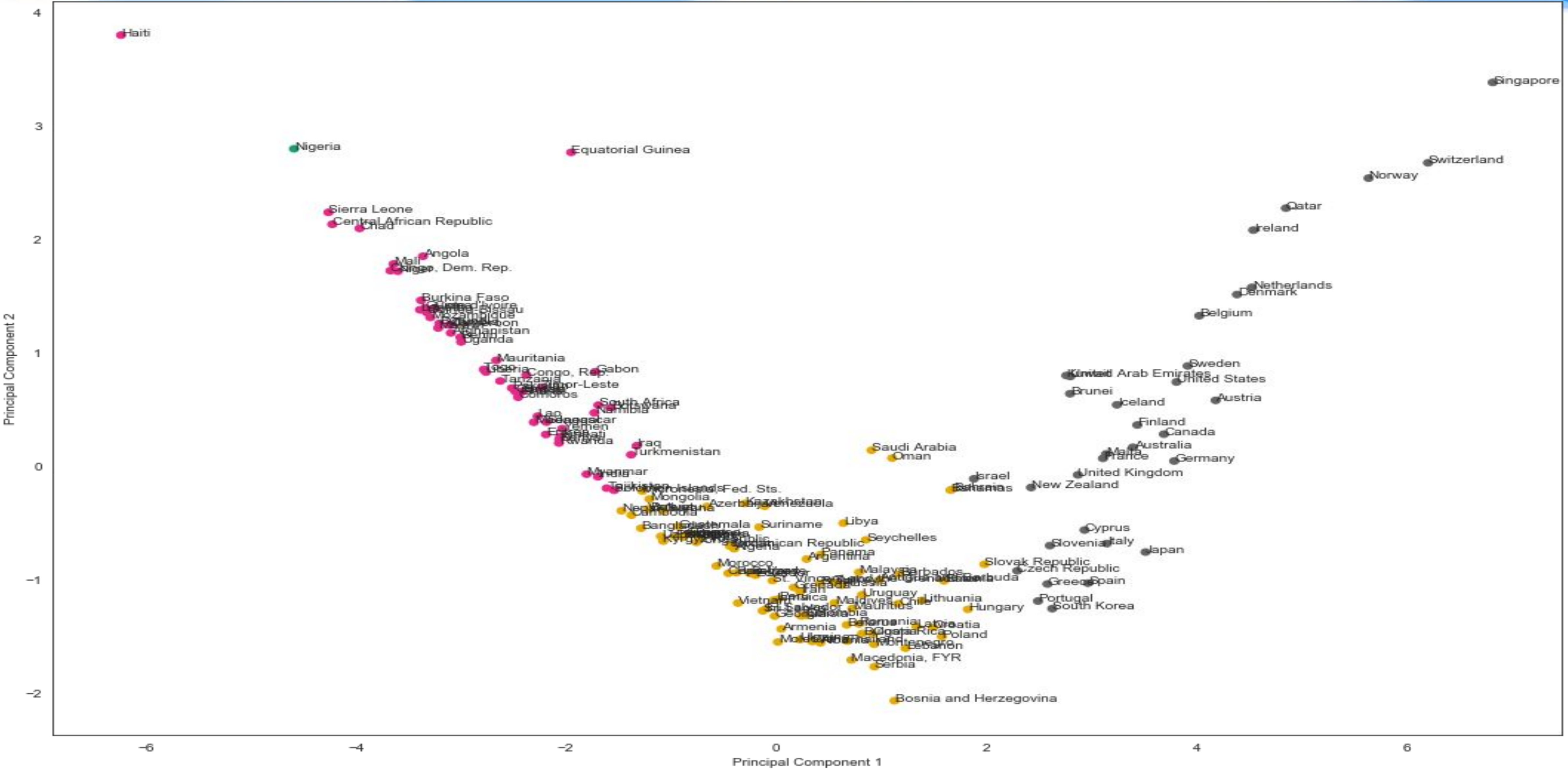
- Clusterid : 2

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

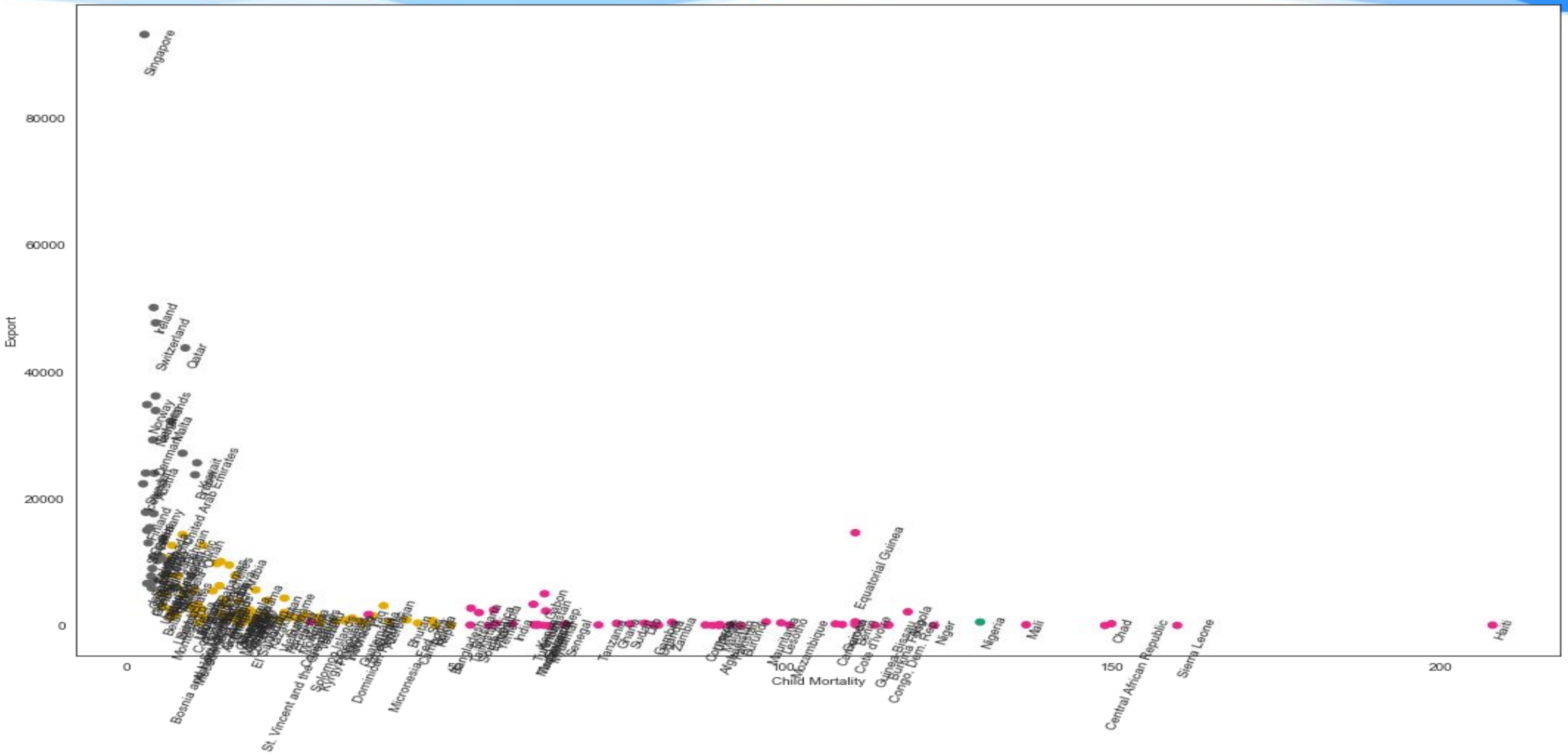
- Clusterid : 3 Luxembourg

- Clusterid : 4 Nigeria

Clustering Using PCA1 & PCA2 Data



Clustering Using Actual Data



Final List of Countries

1. Haiti
2. Sierra Leone
3. Central African Republic
4. Chad
5. Mali

Hopkin Score

Hopkin Score of the created cluster is .854

It means the data in the dataset has high tendency of creating strong, cohesive clusters

Prediction using PCA

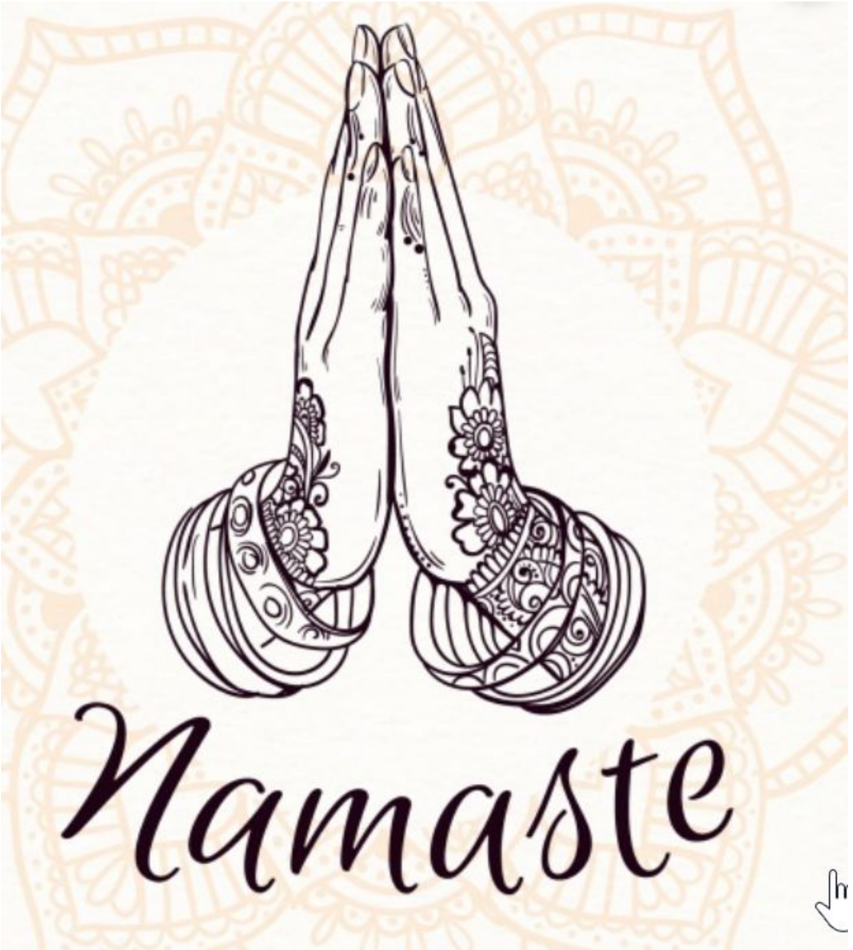
Using Hierarchical Clustering

- Accuracy on Train Dataset : 0.98
- Accuracy on Test Dataset : 0.98

Using KMeans Clustering

- Accuracy on Train Dataset : 0.95
- Accuracy on Test Dataset : 0.90

Thank you


Namaste