# Minor Project PRML
# Report

## Team:

Ankur Yadav (B21CS011)
Adarsh Raj Shrivastava (B21AI003)

## Objective :

To categorize the countries using socio-economic and health factors that determine the overall development of the country.

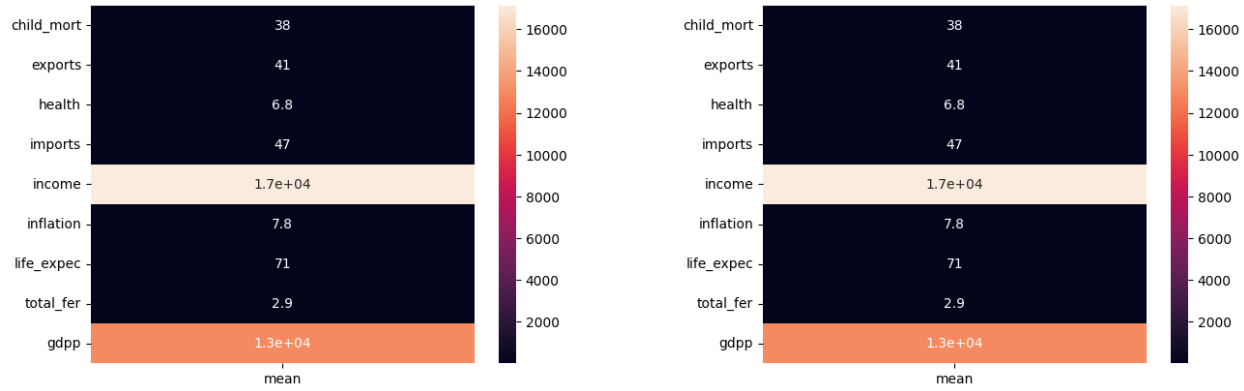
## Importing Dataset :

Imported the country dataset and data description

## Dataset Analysis :

The dataset contains a total 10 features and 167 countries data. There are no null values and all columns are either float values or int values except the first country name column. Then we dropped the country column as it is of no use in clustering.
Following is the data description :

| | child_mort | exports | health | imports | income | inflation | life_expec | total_fer | gdpp |
|---|---|---|---|---|---|---|---|---|---|
| count | 167.000000 | 167.000000 | 167.000000 | 167.000000 | 167.000000 | 167.000000 | 167.000000 | 167.000000 | 167.000000 |
| mean | 38.270060 | 41.108976 | 6.815689 | 46.890215 | 17144.688623 | 7.781832 | 70.555689 | 2.947964 | 12964.155689 |
| std | 40.328931 | 27.412010 | 2.746837 | 24.209589 | 19278.067698 | 10.570704 | 8.893172 | 1.513848 | 18328.704809 |
| min | 2.600000 | 0.109000 | 1.810000 | 0.065900 | 609.000000 | -4.210000 | 32.100000 | 1.150000 | 231.000000 |
| 25% | 8.250000 | 23.800000 | 4.920000 | 30.200000 | 3355.000000 | 1.810000 | 65.300000 | 1.795000 | 1330.000000 |
| 50% | 19.300000 | 35.000000 | 6.320000 | 43.300000 | 9960.000000 | 5.390000 | 73.100000 | 2.410000 | 4660.000000 |
| 75% | 62.100000 | 51.350000 | 8.600000 | 58.750000 | 22800.000000 | 10.750000 | 76.800000 | 3.880000 | 14050.000000 |
| max | 208.000000 | 200.000000 | 17.900000 | 174.000000 | 125000.000000 | 104.000000 | 82.800000 | 7.490000 | 105000.000000 |




Here the Average and Standard Deviation of ‘income’ and ‘gdp’ is very high hence the data needs to be scaled.

Some observations from means and std:

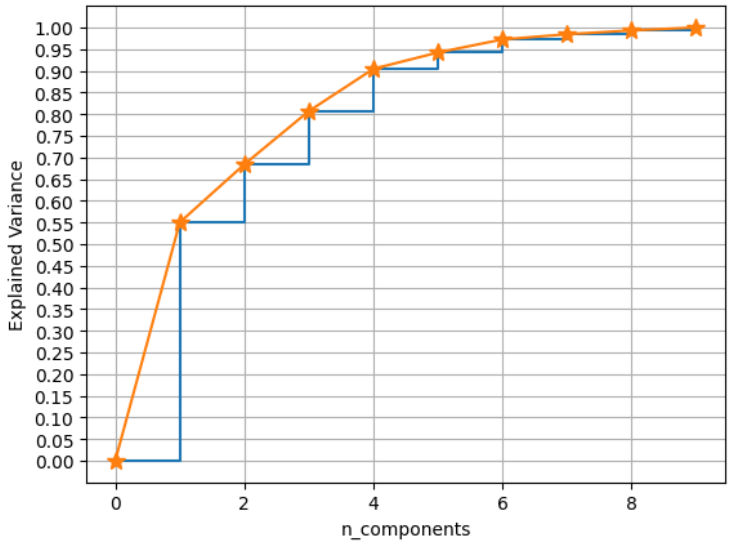
1. Child mortality rate (per 1000 births) ranges from 2.6 to 208 with a mean of 38.27 and a standard deviation of 40.33.
2. Exports of goods and services (% of GDP) ranges from 0.109% to 200% with a mean of 41.11 and a standard deviation of 27.41.
3. Health expenditure per capita ranges from 1.81 to 17.9 with a mean of 6.82 and a standard deviation of 2.75.
4. Imports of goods and services (% of GDP) ranges from 0.066% to 174% with a mean of 46.89 and a standard deviation of 24.21.
5. Income per capita ranges from 609 to 125,000 with a mean of 17,144.69 and a standard deviation of 19,278.07.
6. Inflation, measured by the annual growth rate of the Consumer Price Index (CPI), ranges from -4.21% to 104% with a mean of 7.78 and a standard deviation of 10.57.
7. Life expectancy at birth (years) ranges from 32.1 to 82.8 with a mean of 70.56 and a standard deviation of 8.89.
8. Total fertility rate (children born per woman) ranges from 1.15 to 7.49 with a mean of 2.95 and a standard deviation of 1.51.
9. GDP per capita ranges from 231 to 105,000 with a mean of 12,964.16 and a standard deviation of 18,328.70.

| | min | max | avg |
|---|---|---|---|
| child_mort | Iceland(2.6) | Haiti(208.0) | 38.2701 |
| exports | Myanmar(0.109) | Singapore(200.0) | 41.109 |
| health | Qatar(1.81) | United States(17.9) | 6.81569 |
| imports | Myanmar(0.0659) | Singapore(174.0) | 46.8902 |
| income | Congo, Dem. Rep.(609) | Qatar(125000) | 17144.7 |
| inflation | Seychelles(-4.21) | Nigeria(104.0) | 7.78183 |
| life_expec | Haiti(32.1) | Japan(82.8) | 70.5557 |
| total_fer | Singapore(1.15) | Niger(7.49) | 2.94796 |
| gdpp | Burundi(231) | Luxembourg(105000) | 12964.2 |

## Feature Selection:

### Method 1: Using PCA

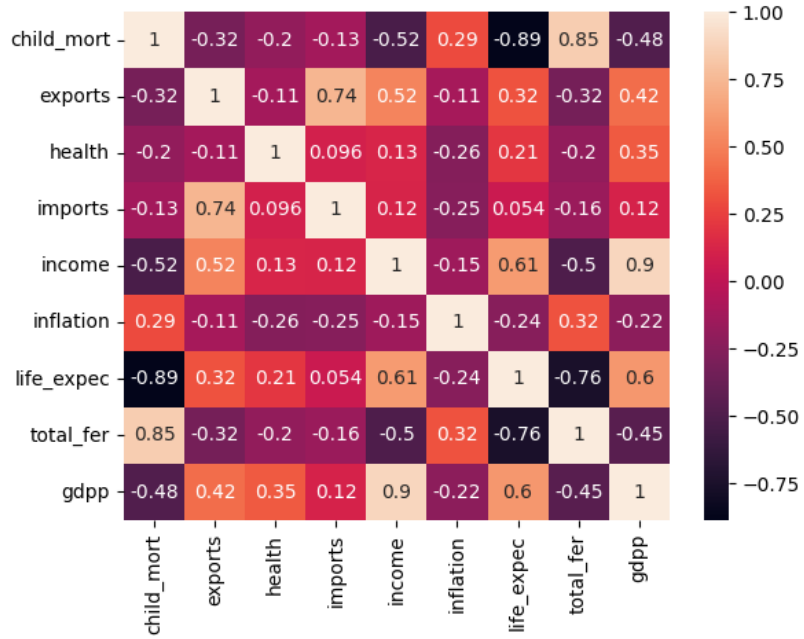
First calculated the feature wise variance.




Here variance greater than 0.95 comes for 5 or more components. So the optimal value of components in PCA is 5. Then we transformed the scaled_data(Using MinMaxScaler) using PCA into 5 components.

## Method 2: Using Grouping

Correlation matrix of features heat map :



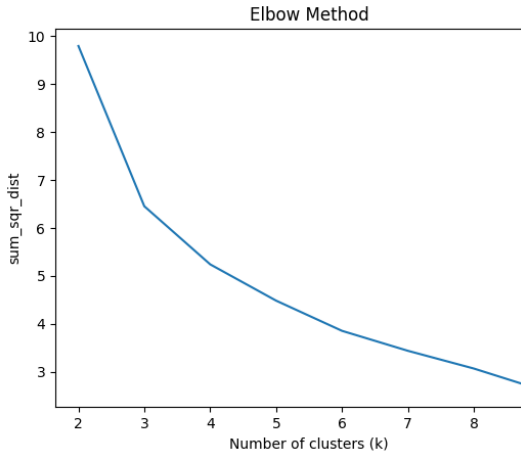

From the heat map we can see that import and export are highly correlated(0.74). So I combined them into one feature that is trade. Similarly I made 4 groups of features.

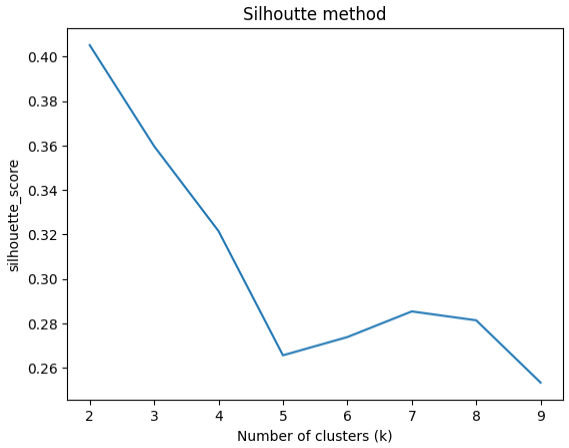
Imports + Exports = Trade
Income + GDP = Economy
Child Mortality + Life expectancy + Total_Fer = Mortality
Inflation = inflation
Health = Health

# Applying Clustering Model :

## 1. K-means :

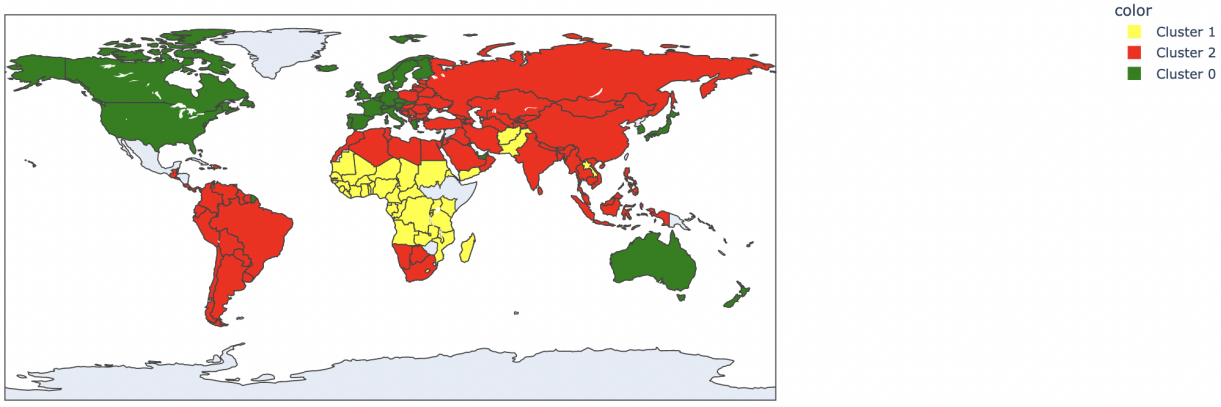
- <u>For Group_data</u> :
  First we found the optimal value of k using the elbow method and silhouette method.




Here we can see that from both the optimal value k comes to be 3.

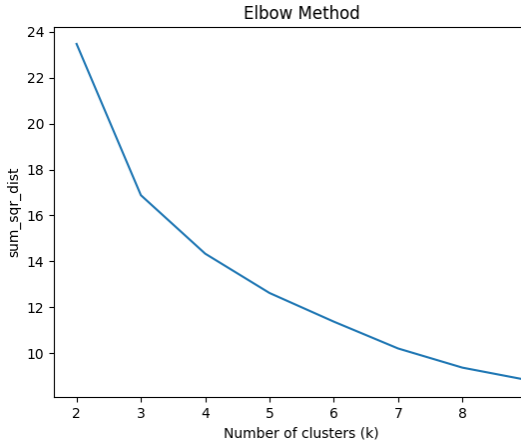
Now we trained the K Means for k=3. Following are the clusters on the world map.

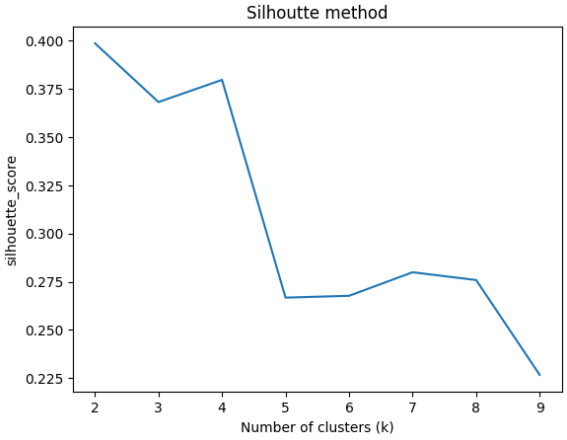



Here we can see that:

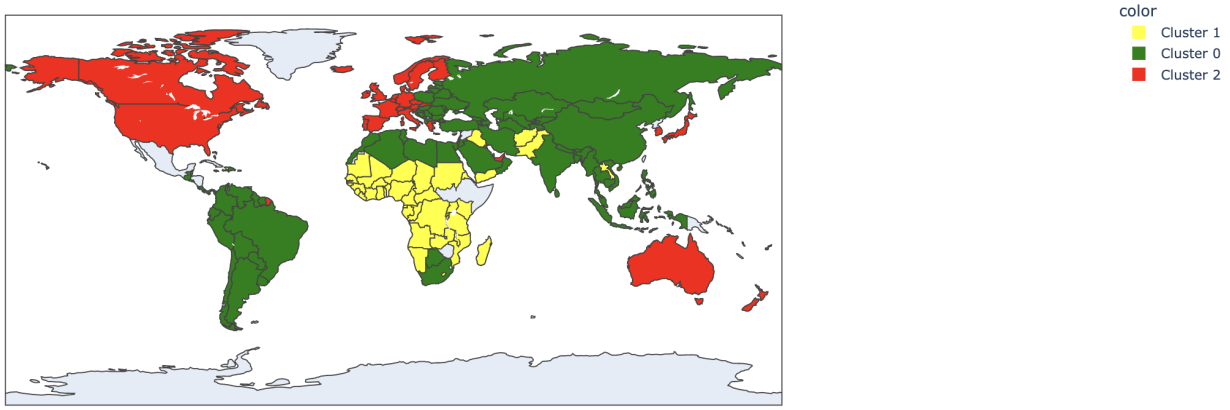
In cluster 0 : USA, Australia, Canada, Norway, France, Belgium, Tokyo etc are in one cluster
In cluster 1 : Most of the African countries
In cluster 2 : India, Russia, China, Mongolia etc are in one cluster

- For PCA_data :




Optimal k comes out to be 3, then we trained data for k = 3.

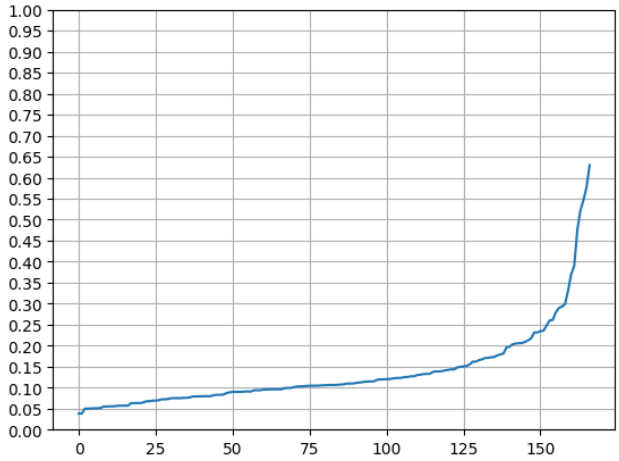



From above world map colour graph we can say that,

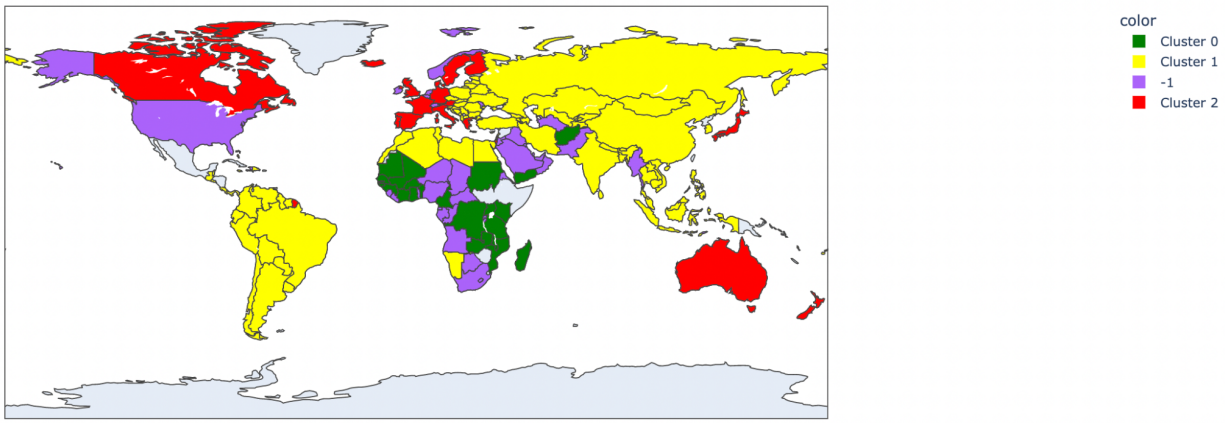
In cluster 0 : USA, Australia, Canada, Norway, France etc are in one cluster
In cluster 1 : Most of the African countries
In cluster 2 : India, Russia, China, Mongolia etc are in one cluster

## 2. DBSCAN :

- For PCA data:



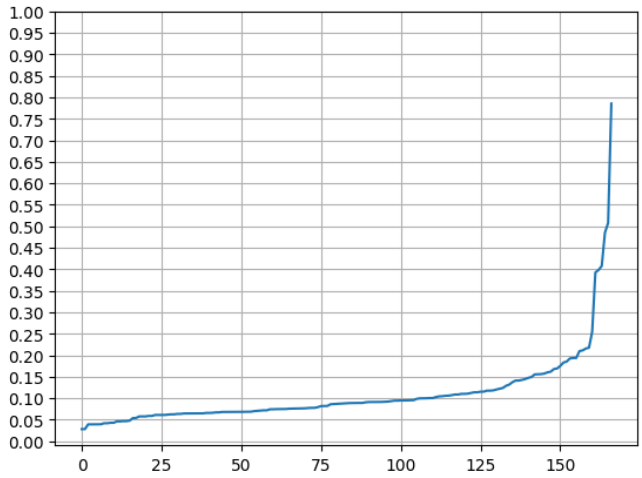

Knee of the plot is at eps = 0.2. Therefore optimal eps = 0.2.



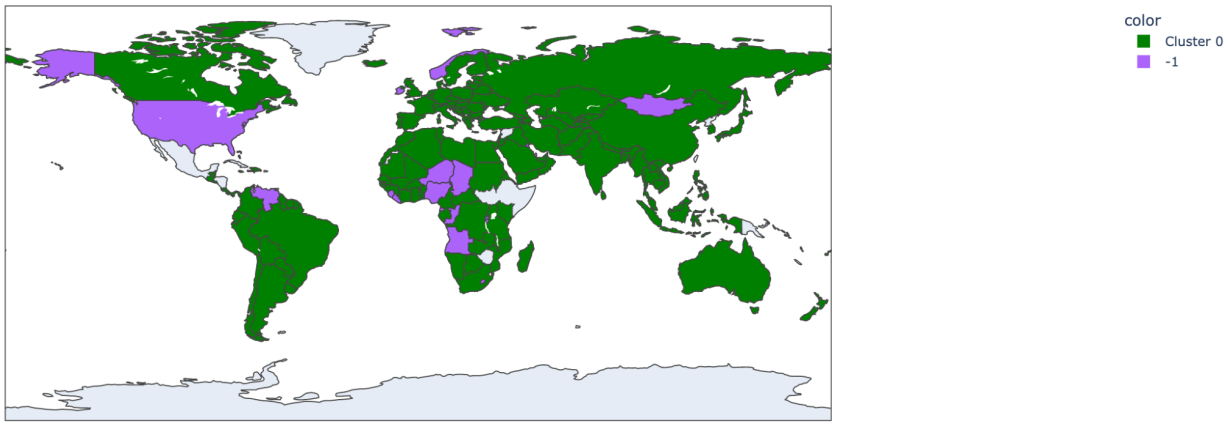

DBSCAN made four clusters:

1. USA, some african countries are treated as outliers
2. Canada, Australia, Spain, Sweden etc are in cluster 2
3. Most of Asian countries and South American countries are in cluster 1
4. Mali, Tanzania, Kenya etc are some countries in cluster 0

- For Group data :




Knee point comes out to be at eps = 0.2, so optimal eps comes out to be eps = 0.2

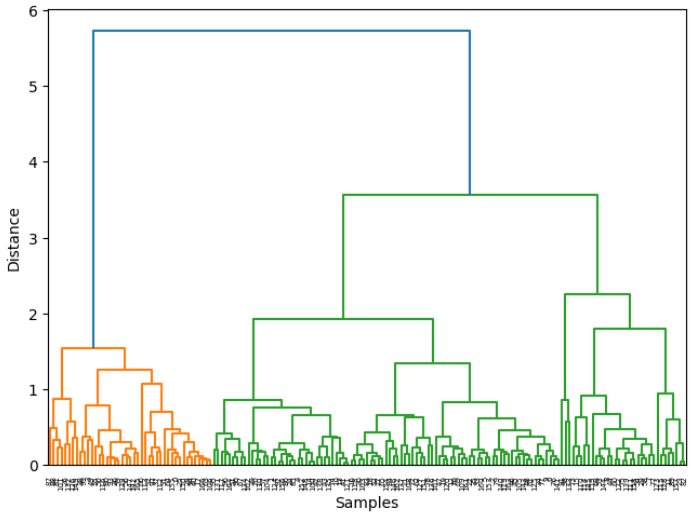


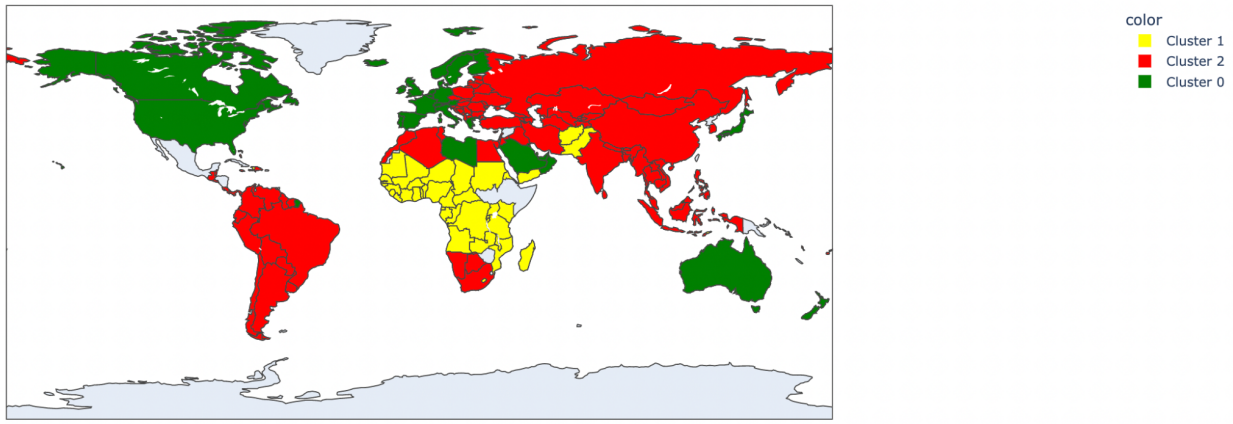

Here, most of the countries are in one cluster, while others are considered as outliers.

Countries that have values that are much higher or lower than the majority of other countries in a particular variable, such as GDP or child mortality rate are considered as outliers.

## 3. Hierarchical :

- For PCA data :

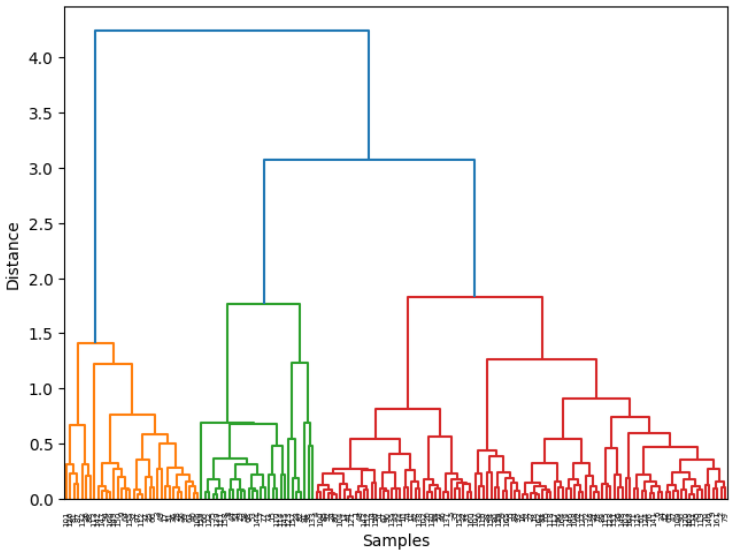


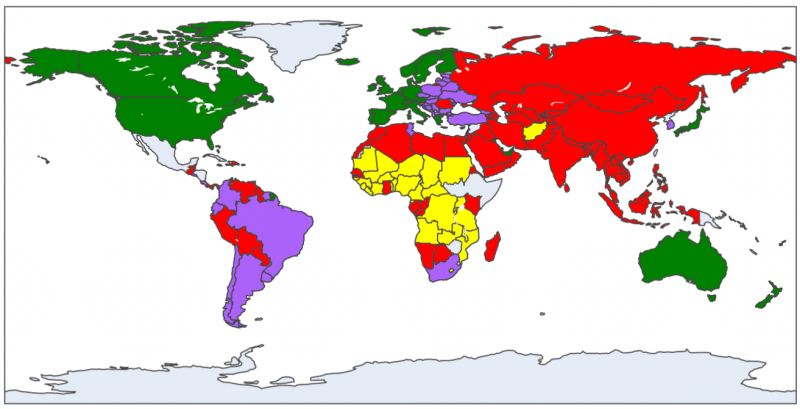




From the above map we can see that

1. USA, Canada, Australia, France etc are in cluster 0
2. Most of the african countries are in cluster 1
3. Most of the Asian countries like India, China, Russia are in cluster 2

- For Group Data :




From the above world map we can see that 4 clusters are made,

1. Countries like USA, Canada, Australia, France etc. are in cluster 0
2. Almost all Asian countries are in cluster 2
3. Some South American countries are in cluster 3.
4. Most of the African countries like Sudan, Chad, Niger etc are in cluster 1.

## Final comparison :

| | PCA Data | Group Data |
|---|---|---|
| KMeans | According to me K means making very good clusters for both PCA and Group data. It is grouping the countries like Japan, USA, most of the European countries, Australia, Canada and New Zealand into one cluster, which we think should actually be in one cluster as they are all developed nations.<br><br>In cluster 2, countries like all South American countries, most of the Asian countries (except Pakistan and Afghanistan), and some of the African countries are grouped. This makes sense as all these countries are developing countries.<br><br>In cluster 3, the algorithm grouped most of the African countries. This also makes sense as these countries are underdeveloped and they lag behind in GDP, infrastructure and health.<br><br>**For grouped data**<br>Cluster 0 : Developed Countries (Green)<br>Cluster 2 : Developing Countries (Red)<br>Cluster 1 : Underdeveloped Countries (Yellow)<br><br>**For PCA data**<br>Cluster 0 : Developing Countries (Green)<br>Cluster 2 : Developed Countries (Red)<br>Cluster 1 : Underdeveloped Countries (Yellow) | |
| DBSCAN | Made 3 clusters of whole data European countries, Australia, and Japan are grouped in 1 cluster.<br><br>All South American and most Asian countries are clustered in 1 group.<br><br>In 3 rd cluster only some African countries are grouped.<br><br>Many countries are classified as outlier<br><br>Seems not a good clustering model for the data | Grouped the whole data in 1 cluster and remaining countries as outliers.<br><br>Not a good model for this data |

| | |
|---|---|
| Hierarchical | The hierarchical model is making almost the same clusters for PCA and grouped data. It is giving almost the same results as K-Means clustering.<br><br>For grouped data, 4 clusters were made, while for PCA data, only 3 were made.<br><br>**For grouped data**<br>Cluster 0 : Fully Developed Countries (Green)<br>Cluster 2 : Developing Countries (Red)<br>Cluster 3 : Near Developed(Purple)<br>Cluster 1 : Underdeveloped Countries (Yellow)<br><br>**For PCA data**<br>Cluster 0 : Developing Countries (Red)<br>Cluster 2 : Developed Countries (Green)<br>Cluster 1 : Underdeveloped Countries (Yellow) |