Project Report

Team:

Anshul Thakur (B21CS085)
Anupam Verma (B21EE007)

Link:

https://colab.research.google.com/drive/19ei7FvqRkfcLRQY5Y81Lptm7Ws43jpJk?usp=sharing

Objective :

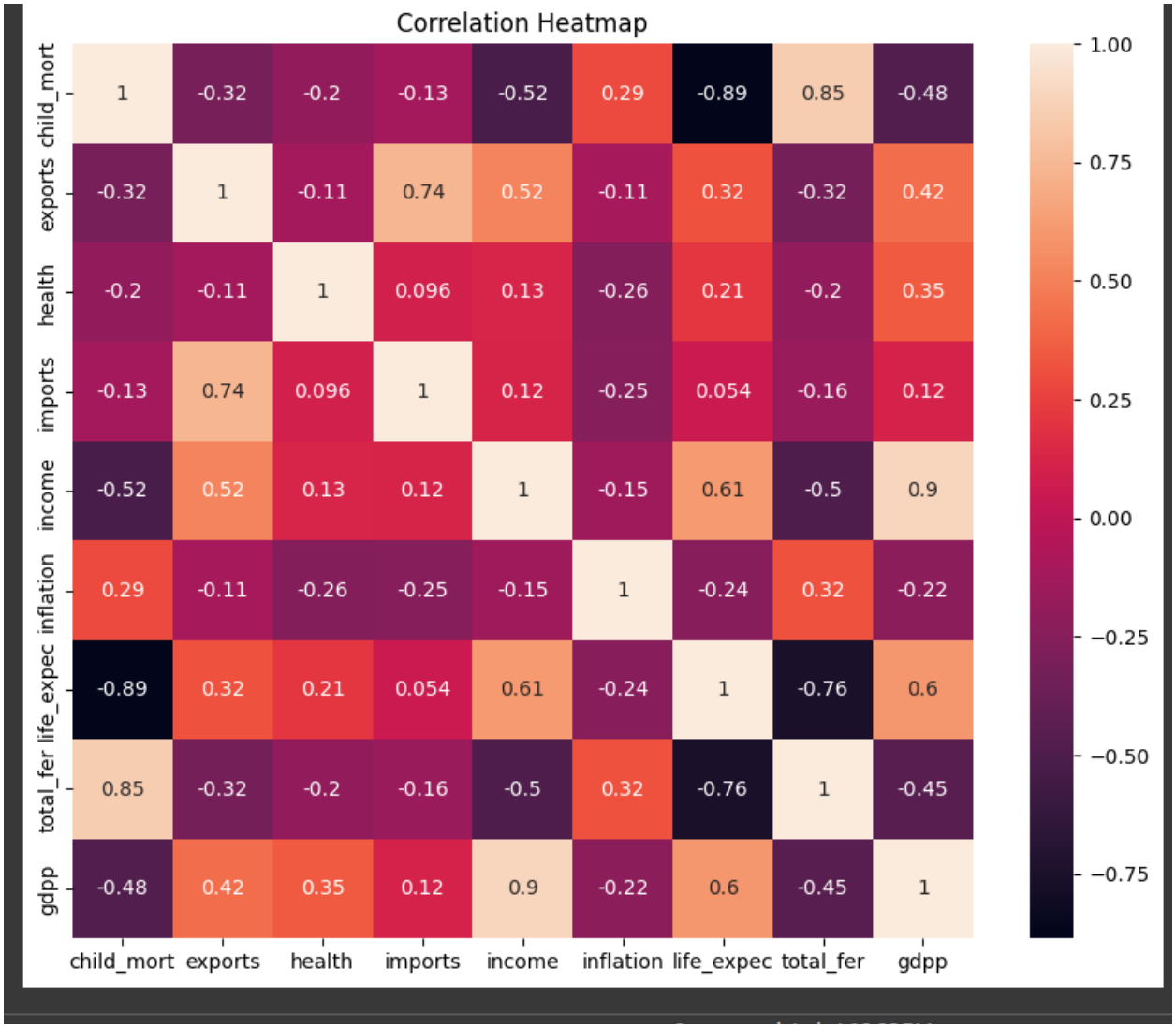
To categorize the countries using socio-economic and health factors that determine the overall development of the country.

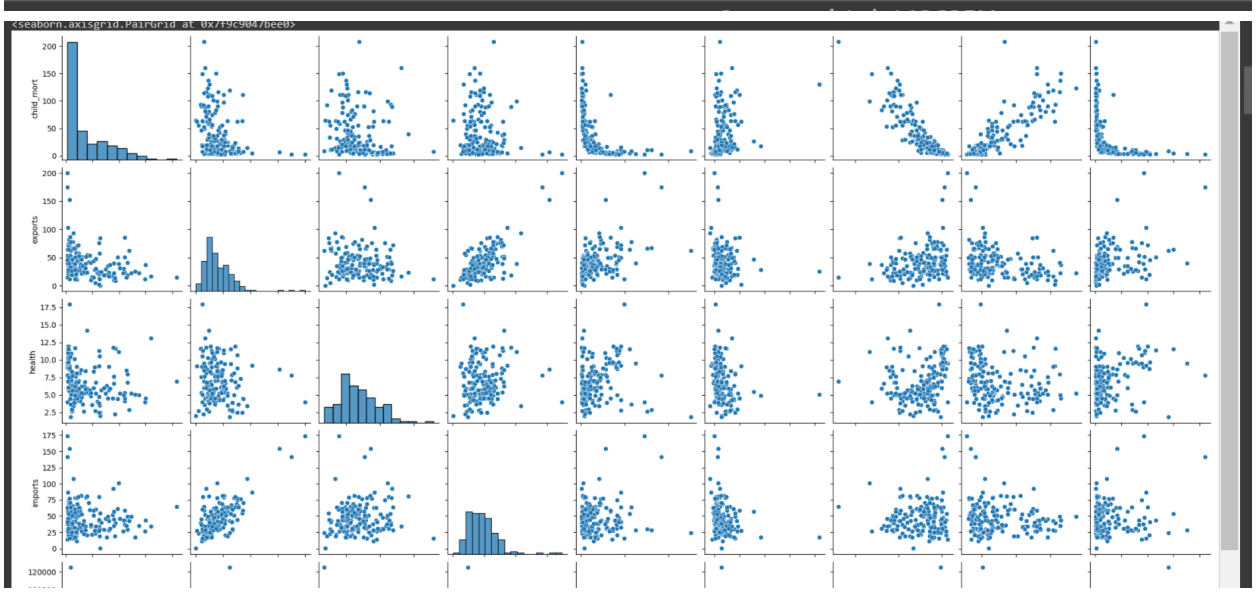
Dataset Analysis :

The dataset contains a total 10 features and 167 countries data. All columns are either float values or int values except the first country name column which is string.

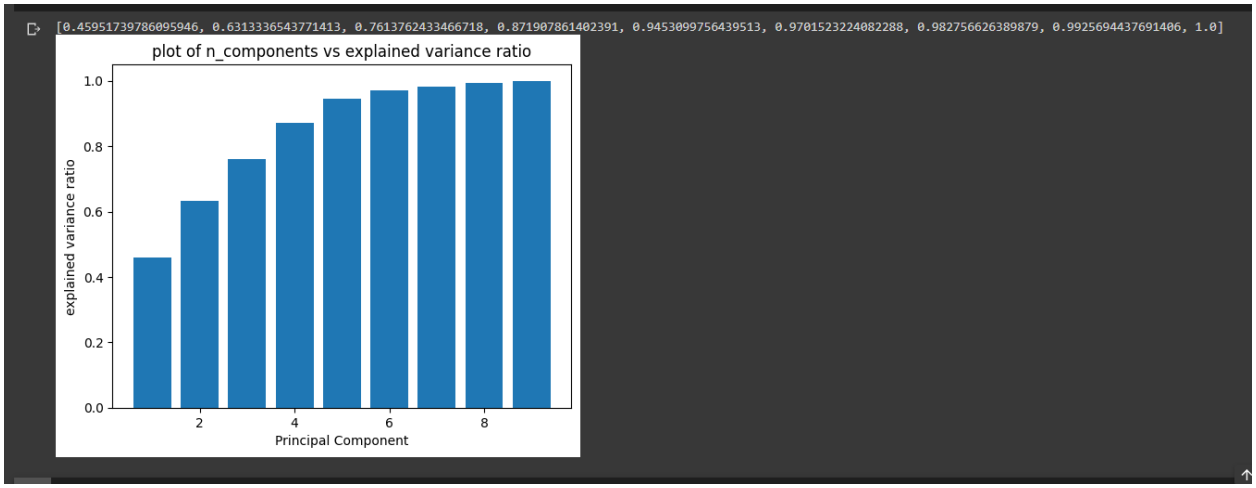
	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
0	Afghanistan	90.2	10.0	7.58	44.9	1610	9.44	56.2	5.82	553
1	Albania	16.6	28.0	6.55	48.6	9930	4.49	76.3	1.65	4090
2	Algeria	27.3	38.4	4.17	31.4	12900	16.10	76.5	2.89	4460
3	Angola	119.0	62.3	2.85	42.9	5900	22.40	60.1	6.16	3530
4	Antigua and Barbuda	10.3	45.5	6.03	58.9	19100	1.44	76.8	2.13	12200
...	...	...	...	...	...	...	...	...	...	...
162	Vanuatu	29.2	46.6	5.25	52.7	2950	2.62	63.0	3.50	2970
163	Venezuela	17.1	28.5	4.91	17.6	16500	45.90	75.4	2.47	13500
164	Vietnam	23.3	72.0	6.84	80.2	4490	12.10	73.1	1.95	1310
165	Yemen	56.3	30.0	5.18	34.4	4480	23.60	67.5	4.67	1310
166	Zambia	83.1	37.0	5.89	30.9	3280	14.00	52.0	5.40	1460

167 rows × 10 columns


Correlation Heatmap
child_mort
exports
health
imports
income
inflation
life_expec
total_fer
gdpp
1 -0.32 -0.2 -0.13 -0.52 0.29 -0.89 0.85 -0.48
-0.32 1 -0.11 0.74 0.52 -0.11 0.32 -0.32 0.42
-0.2 -0.11 1 0.096 0.13 -0.26 0.21 -0.2 0.35
-0.13 0.74 0.096 1 0.12 -0.25 0.054 -0.16 0.12
-0.52 0.52 0.13 0.12 1 -0.15 0.61 -0.5 0.9
0.29 -0.11 -0.26 -0.25 -0.15 1 -0.24 0.32 -0.22
-0.89 0.32 0.21 0.054 0.61 -0.24 1 -0.76 0.6
0.85 -0.32 -0.2 -0.16 -0.5 0.32 -0.76 1 -0.45
-0.48 0.42 0.35 0.12 0.9 -0.22 0.6 -0.45 1
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75


child_mort
exports
health
imports

PCA:

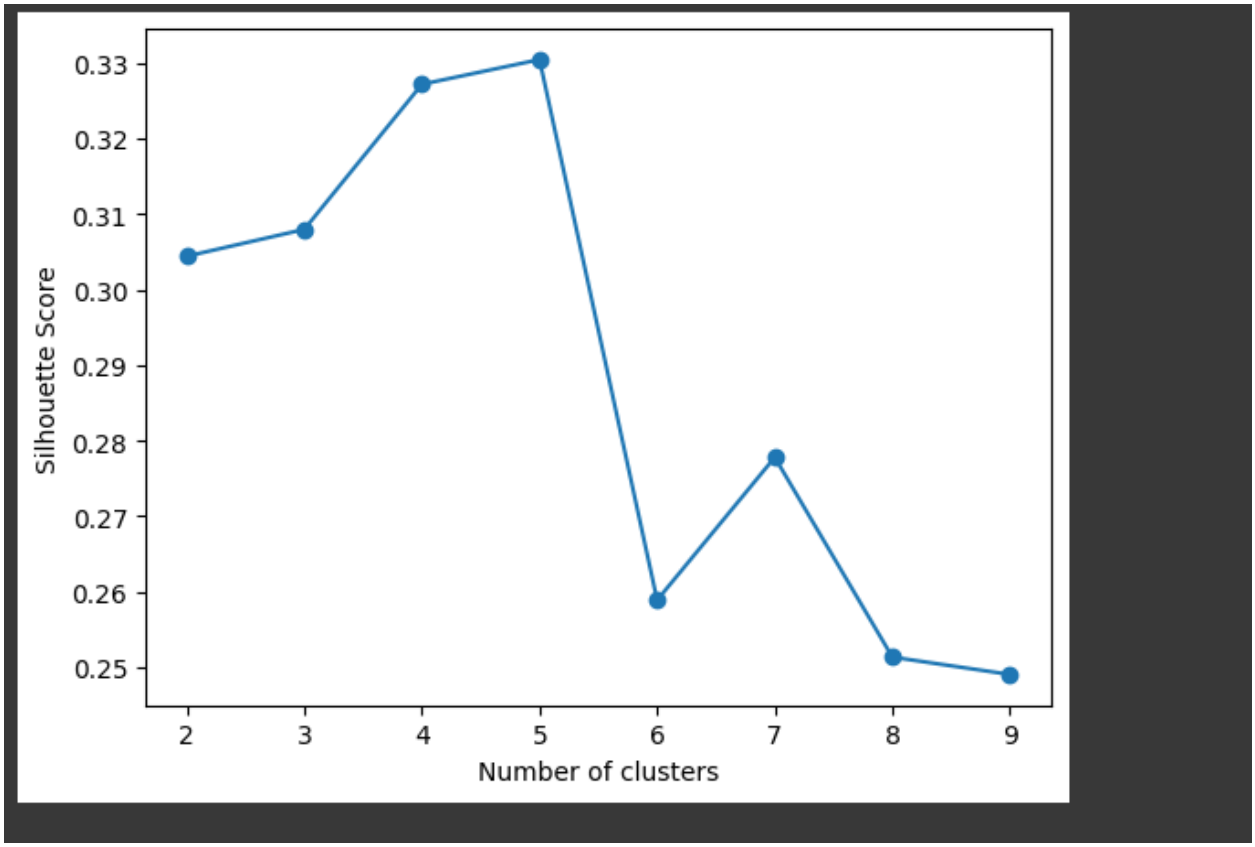


We used PCA to reduce the data's dimension to 5 as the cumulative explained variance is reaching 0.95 for n_components = 5.

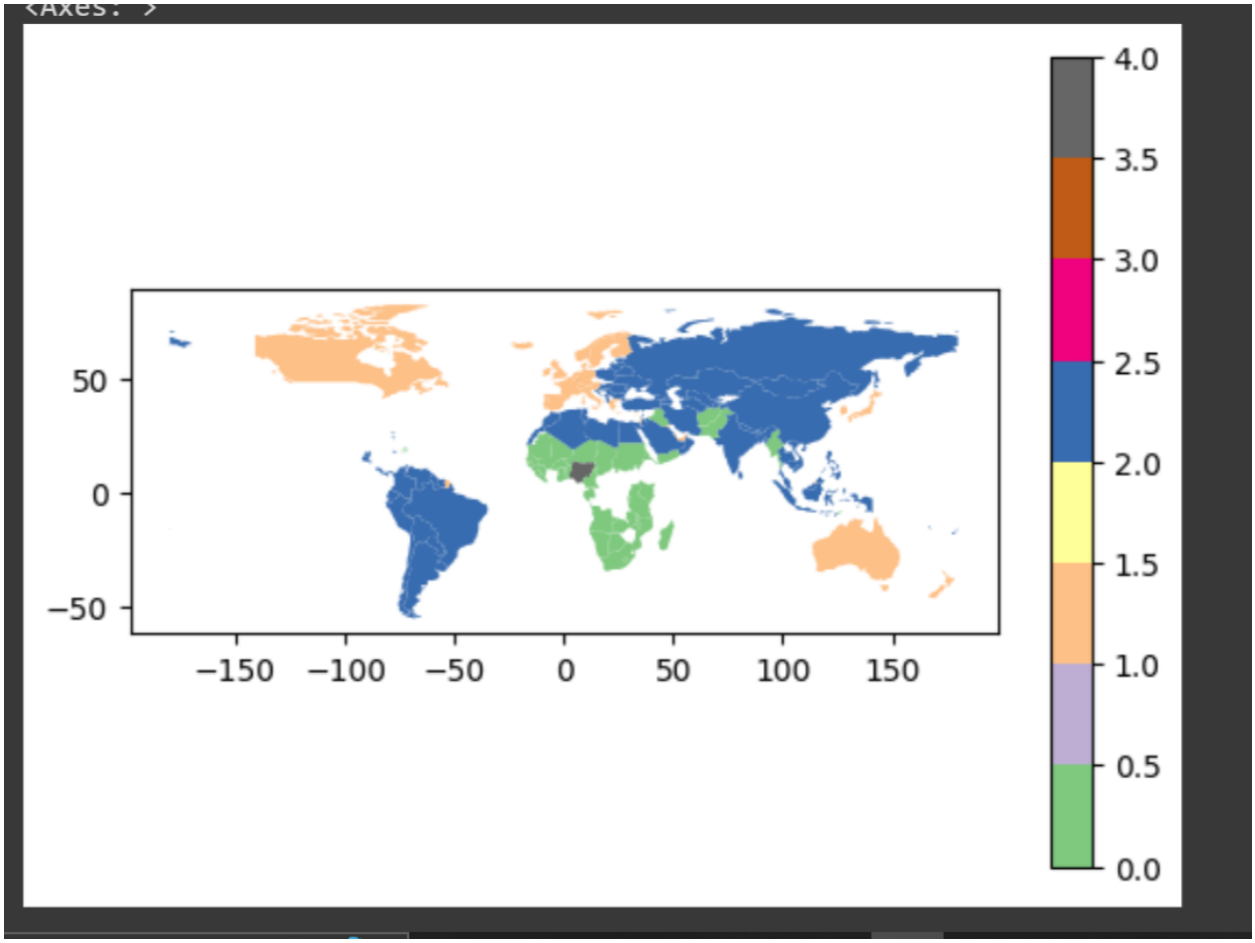
Applying Clustering Model :

1. K-means :

First I found the optimal value of k using the silhouette score.



Here we can see that from both the optimal value k comes to be 5

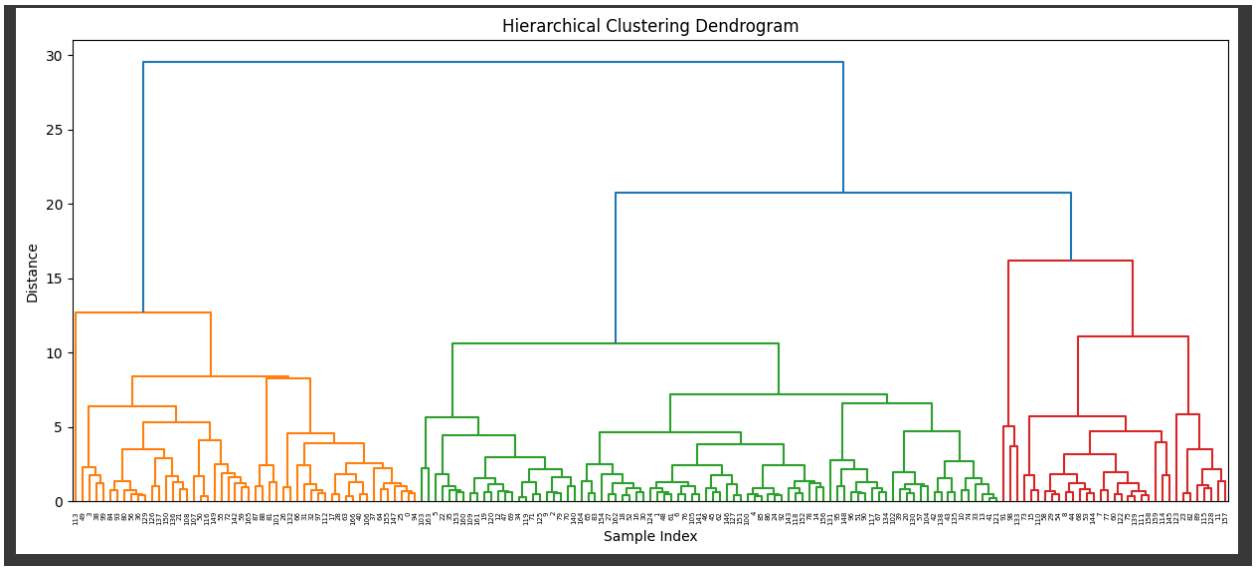


From the above world map we can see that 4 clusters are made,

1. Countries like USA, Canada, Australia, France etc. are in 1st cluster
2. Almost all Asian countries are in the 2nd cluster.
3. Almost all African countries along with pakistan and afghanistan are in the 3rd cluster.
4. One country in Africa is labeled as the fourth cluster

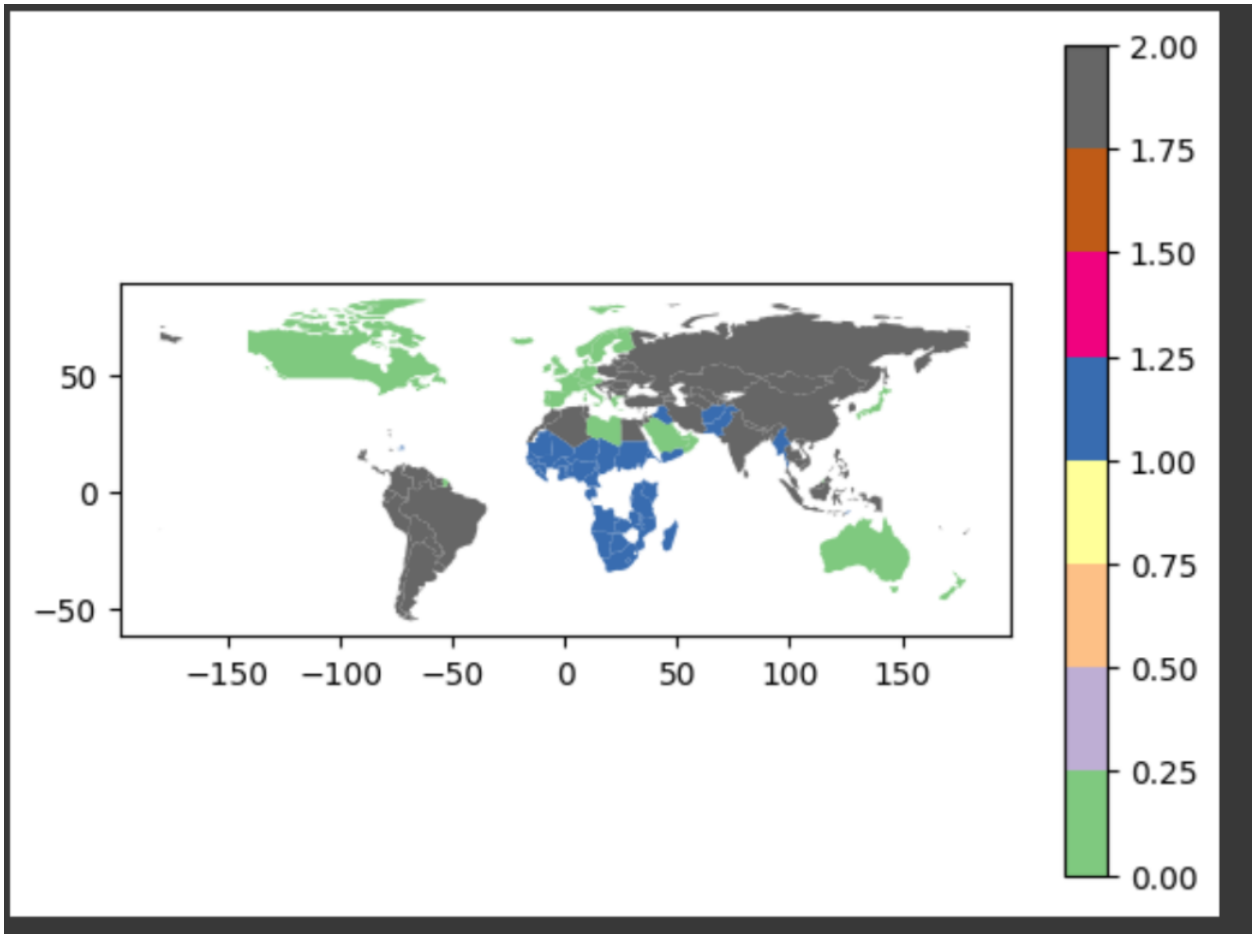
2. Hierarchy Clustering:

   The top-down hierarchical clustering algorithm



From this we find out the optimal number of clusters for hierarchical clustering.

Then we apply agglomerative clustering to the pca data.



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

1. Countries like USA, Canada, Australia, France etc. are in 1st cluster
2. Almost all Asian countries are in the 2nd cluster.
3. African countries along with Pakistan and Afghanistan are in the 3rd cluster.

Final comparison :

1. KMeans Clustering:

K means clustering makes fairly good clusters, as we can see. It places Japa, the United States, Australia, and certain parts of Europe in one cluster; we can see that these constitute the developed countries.

In the second cluster we have major Asian countries (including India) and Russia, and some African countries. The model placed these countries in one group as these are developing countries.

In the third cluster we have most of the African countries which are grouped so as they constitute the underdeveloped countries.

There is a fourth and Fifth cluster which constitute a country each both representing extremely underdeveloped countries.

2. Hierarchical Clustering:

   First using top-down hierarchical clustering we estimated the optimal number of clusters that come out to be three.

   Rest the grouping of countries is similar to that of K Means clustering.