CLUSTERING OF COUNTRIES

SOLON KUMAR DAS

CLUSTERING OF COUNTRIES

BUSINESS UNDERSTANDING

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.

And this is where you come in as a data analyst. Your job is to categorise the countries using some socio-economic and health factors that determine the overall development of the country. Then you need to suggest the countries which the CEO needs to focus on the most.

CLUSTERING OF COUNTRIES

DATA UNDERSTANDING

The Country-Data consists of the following columns:

Column Name	Description
country	Name of the country
child_mort	Death of children under 5 years of age per 1000 live births
exports	Exports of goods and services. Given as %age of the Total GDP
health	Total health spending as %age of Total GDP
imports	Imports of goods and services. Given as %age of the Total GDP
income	Net income per person
inflation	The measurement of the annual growth rate of the Total GDP
life_expec	The average number of years a new born child would live if the current mortality patterns are to remain the same
total_fer	The number of children that would be born to each woman if the current age-fertility rates remain the same.
gdpp	The GDP per capita. Calculated as the Total GDP divided by the total population.

CLUSTERING OF COUNTRIES

PRINCIPAL COMPONENT ANALYSIS

CLUSTERING OF COUNTRIES

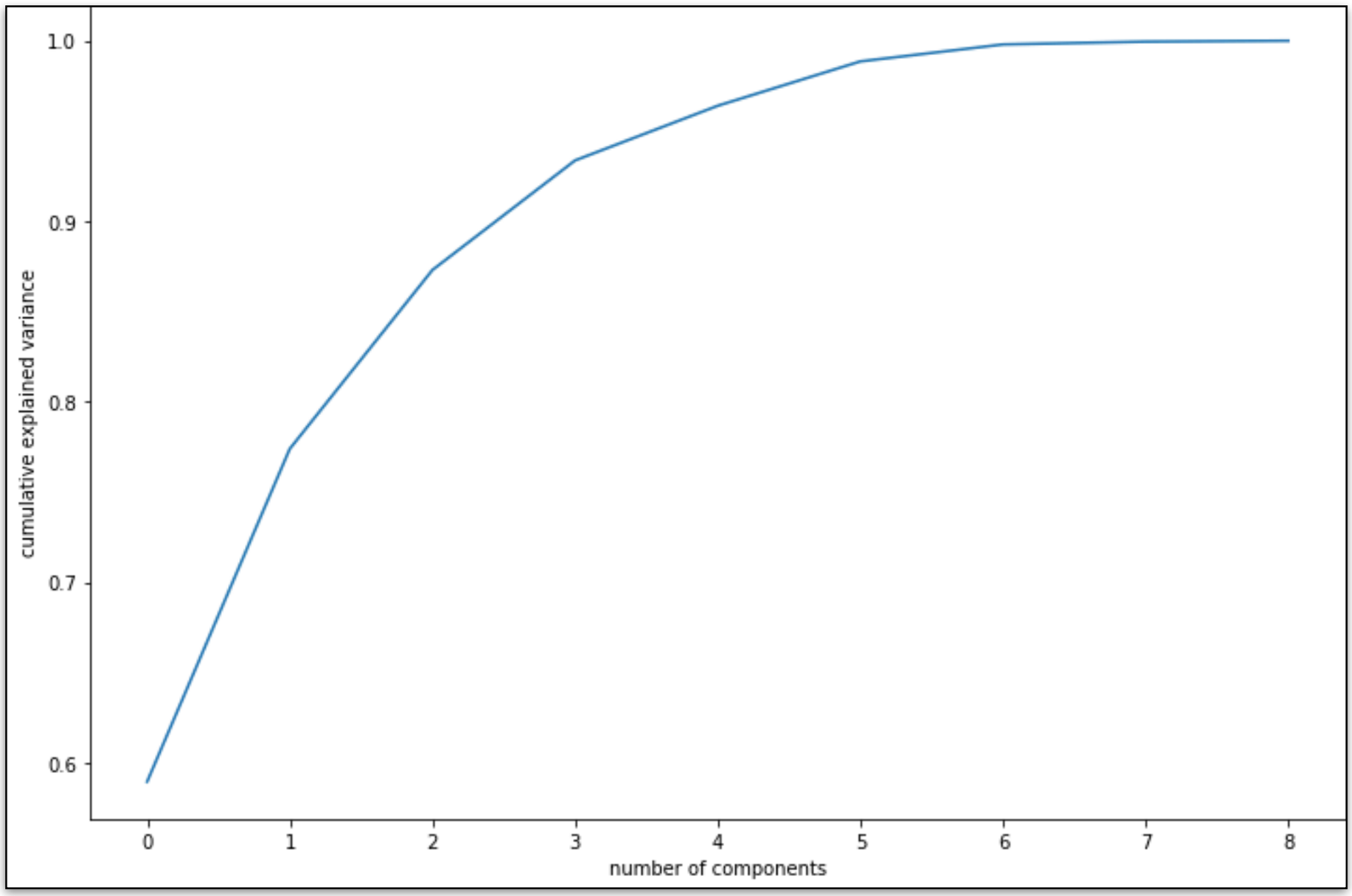
PCA is done to achieve dimentionality reduction, to remove Multicollinearity among the variables and produce a stable set of variables for models to build upon.

Lets have a look at how our correlation data looks before PCA is performed.

We indeed see some correlation within the variables so we have to perform PCA to omit this.

While conducting PCA we plot something called as a **Scree Plot** which tells us what number of principal component explains the most amount of variance in the data.

Scree Plot

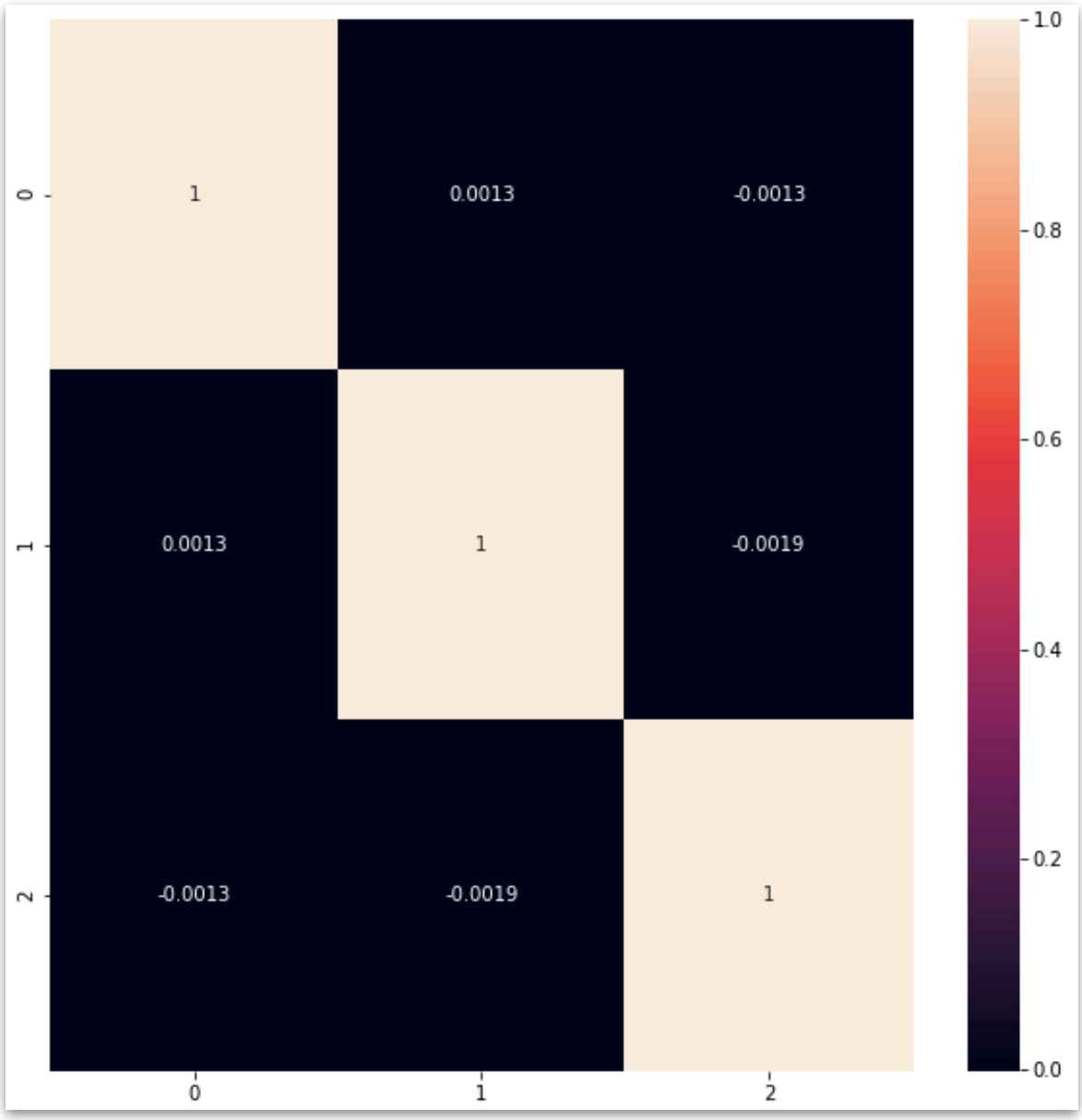


Conclusion:

Clearly over 90% of the data is properly explained by the first 3 principal components.
We shall use these three components for our clustering process.

CLUSTERING OF COUNTRIES

After performing PCA with 3 Components, we look at our correlation map again.



Conclusion : we see that our principal components are completely uncorrelated.

After PCA the data head looks as such:

	Feature	PC1	PC2	PC3
0	child_mort	-0.316392	0.476267	-0.150012
1	exports	0.342887	0.397311	-0.030574
2	health	0.358535	0.155053	-0.075703
3	imports	0.344865	0.370781	-0.072174
4	income	0.380041	0.128384	0.145764
5	inflation	-0.143085	0.221261	0.948419
6	life_expec	0.343857	-0.369820	0.196752
7	total_fer	-0.302842	0.459715	-0.077834
8	gdpp	0.399988	0.200624	0.010339

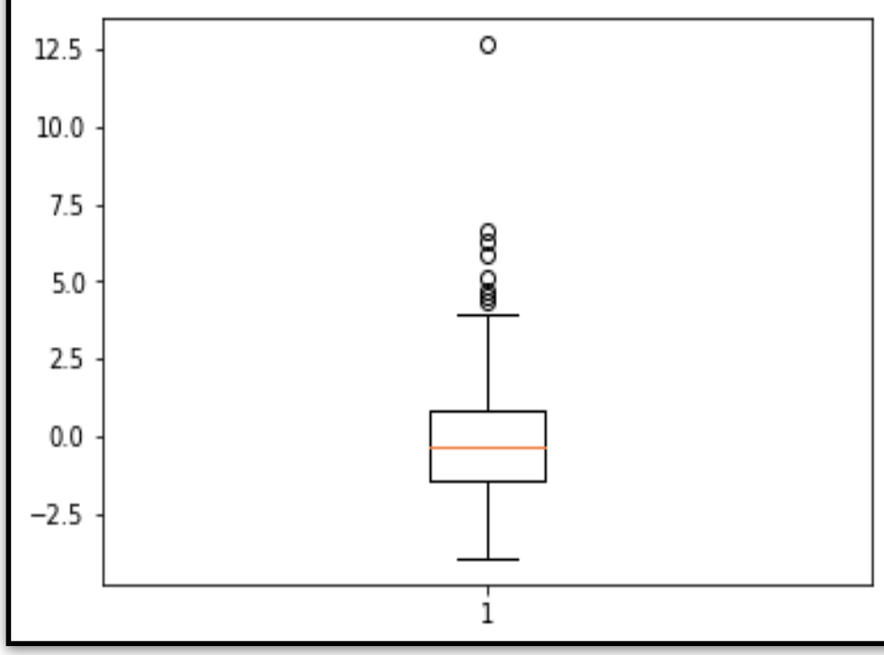
OUTLIER ANALYSIS AND TREATMENT

CLUSTERING OF COUNTRIES

Before proceeding to clustering we should always do a Outlier treatment of the data because presence of outlier in the data can effect clusters badly and may result in bad clustering.
Therefore we will do three outlier treatments for each Principle Component Chosen.

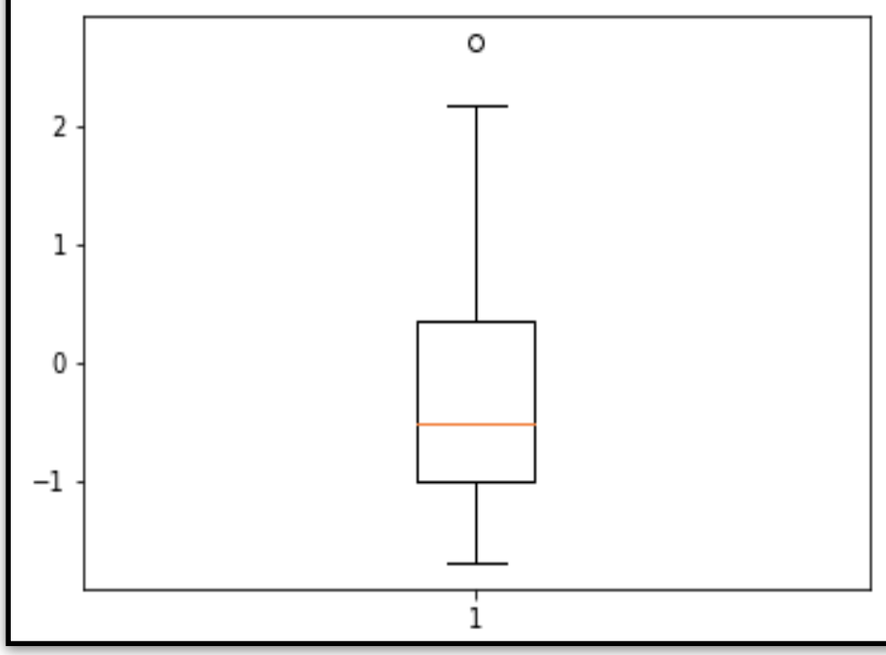
For PC1

```
plt.boxplot(pcs_df2.PC1)
Q1 = pcs_df2.PC1.quantile(0.05)
Q3 = pcs_df2.PC1.quantile(0.95)
IQR = Q3 - Q1
pcs_df2 = pcs_df2[(pcs_df2.PC1 >= Q1) & (pcs_df2.PC1 <= Q3)]
```



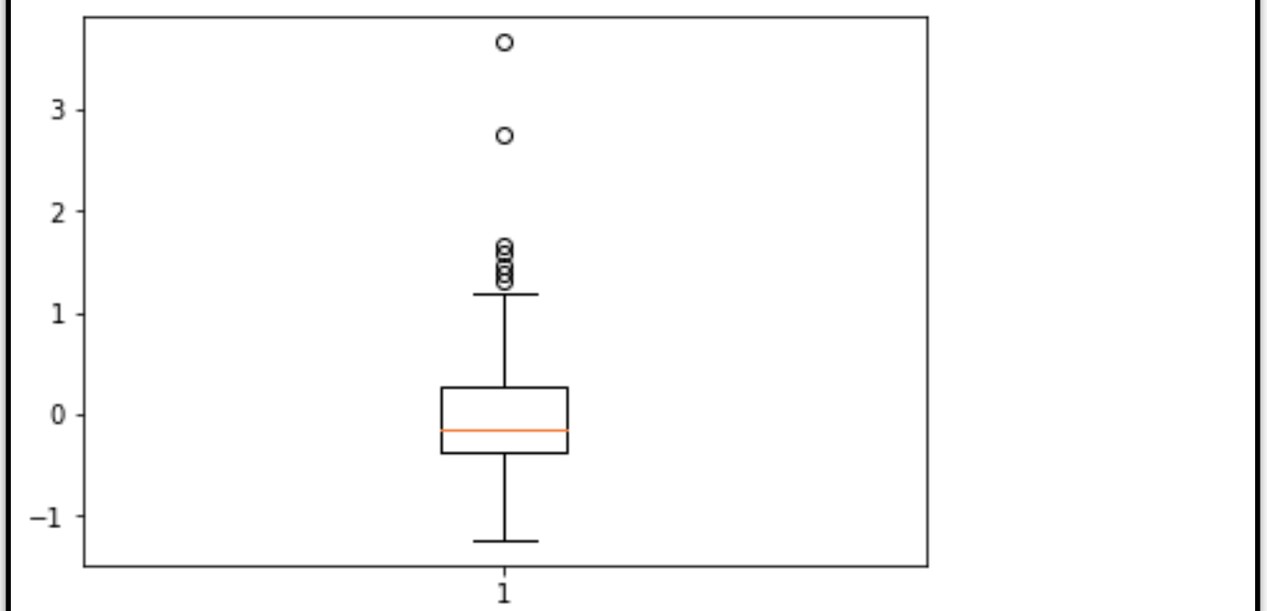
For PC2

```
plt.boxplot(pcs_df2.PC2)
Q1 = pcs_df2.PC2.quantile(0.05)
Q3 = pcs_df2.PC2.quantile(0.95)
IQR = Q3 - Q1
pcs_df2 = pcs_df2[(pcs_df2.PC2 >= Q1) & (pcs_df2.PC2 <= Q3)]
```



For PC3

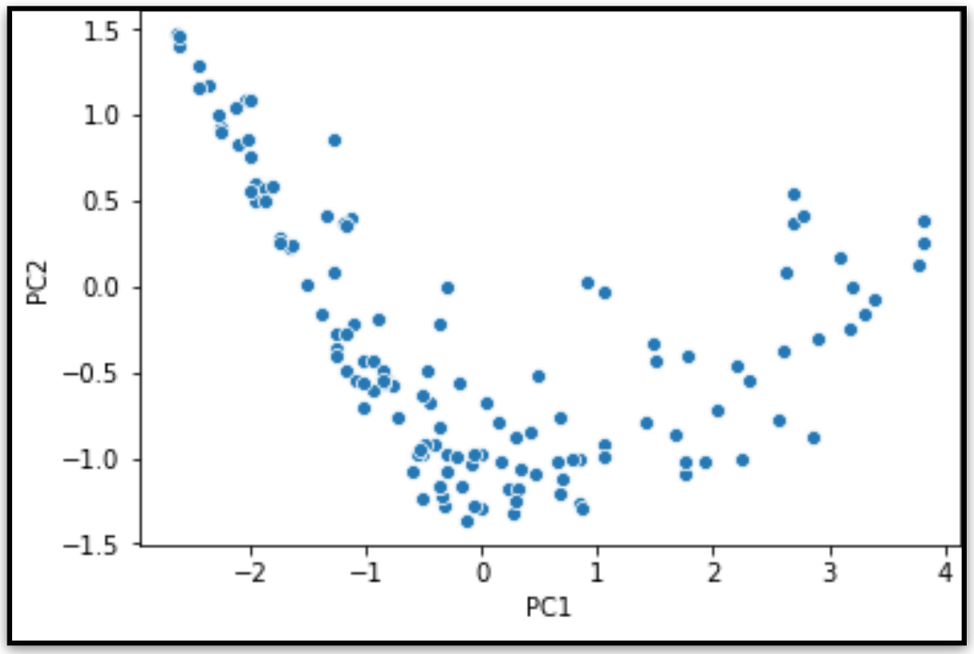
```
plt.boxplot(pcs_df2.PC3)
Q1 = pcs_df2.PC3.quantile(0.05)
Q3 = pcs_df2.PC3.quantile(0.95)
IQR = Q3 - Q1
dat3 = pcs_df2[(pcs_df2.PC3 >= Q1 ) & (pcs_df2.PC3 <= Q3)]
```



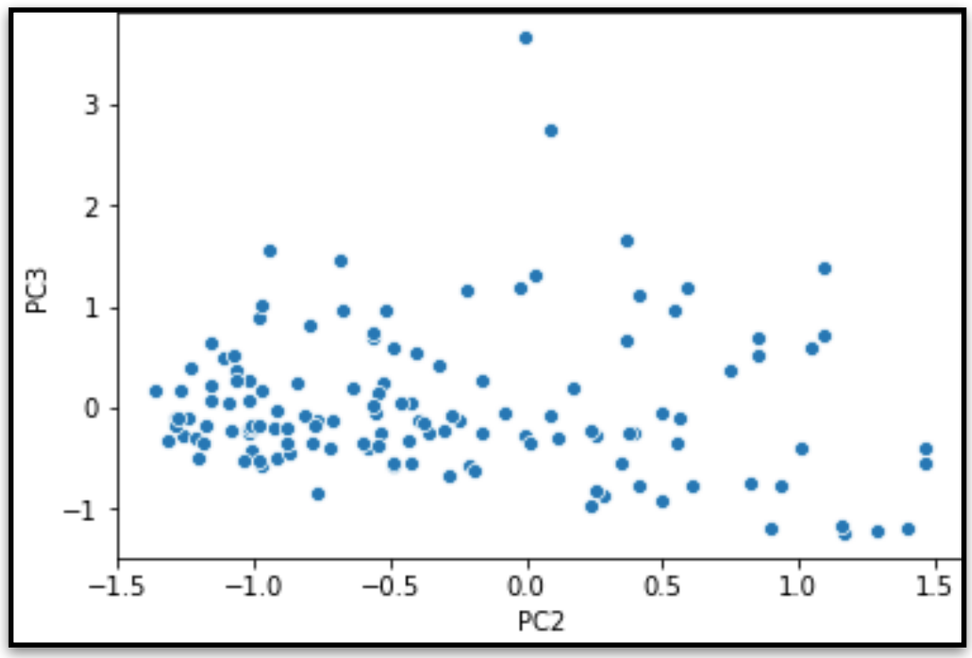
For each Principal Component we have removed the outliers that lie below 5 % and above 95%

CLUSTERING OF COUNTRIES

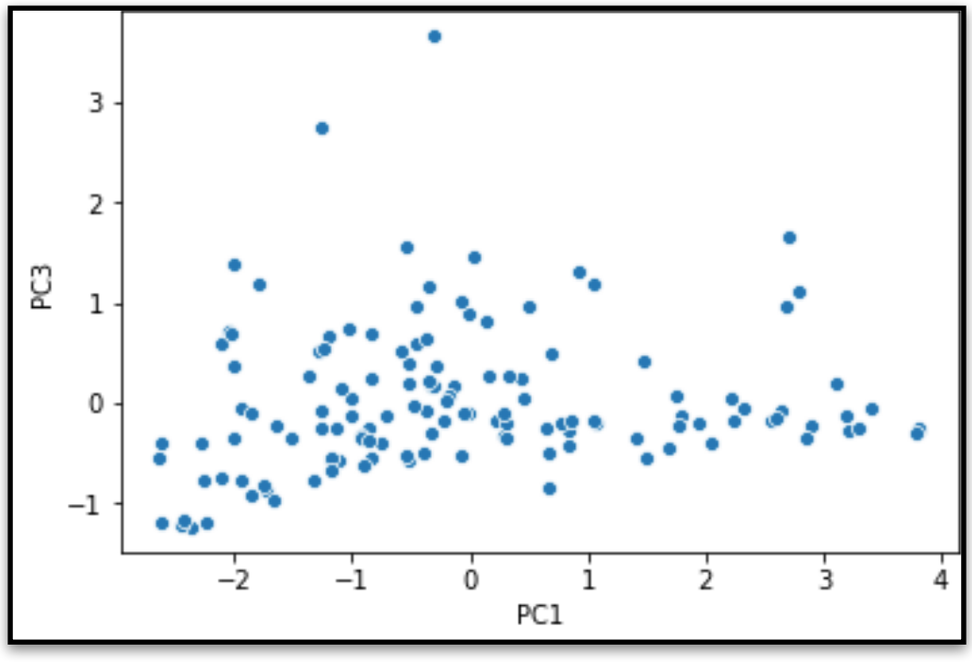
Before we proceed lets have a look at the spread of the data in terms of the principle component.



PC1 vs PC2



PC2 vs PC3



PC1 vs PC3

CLUSTERING OF COUNTRIES

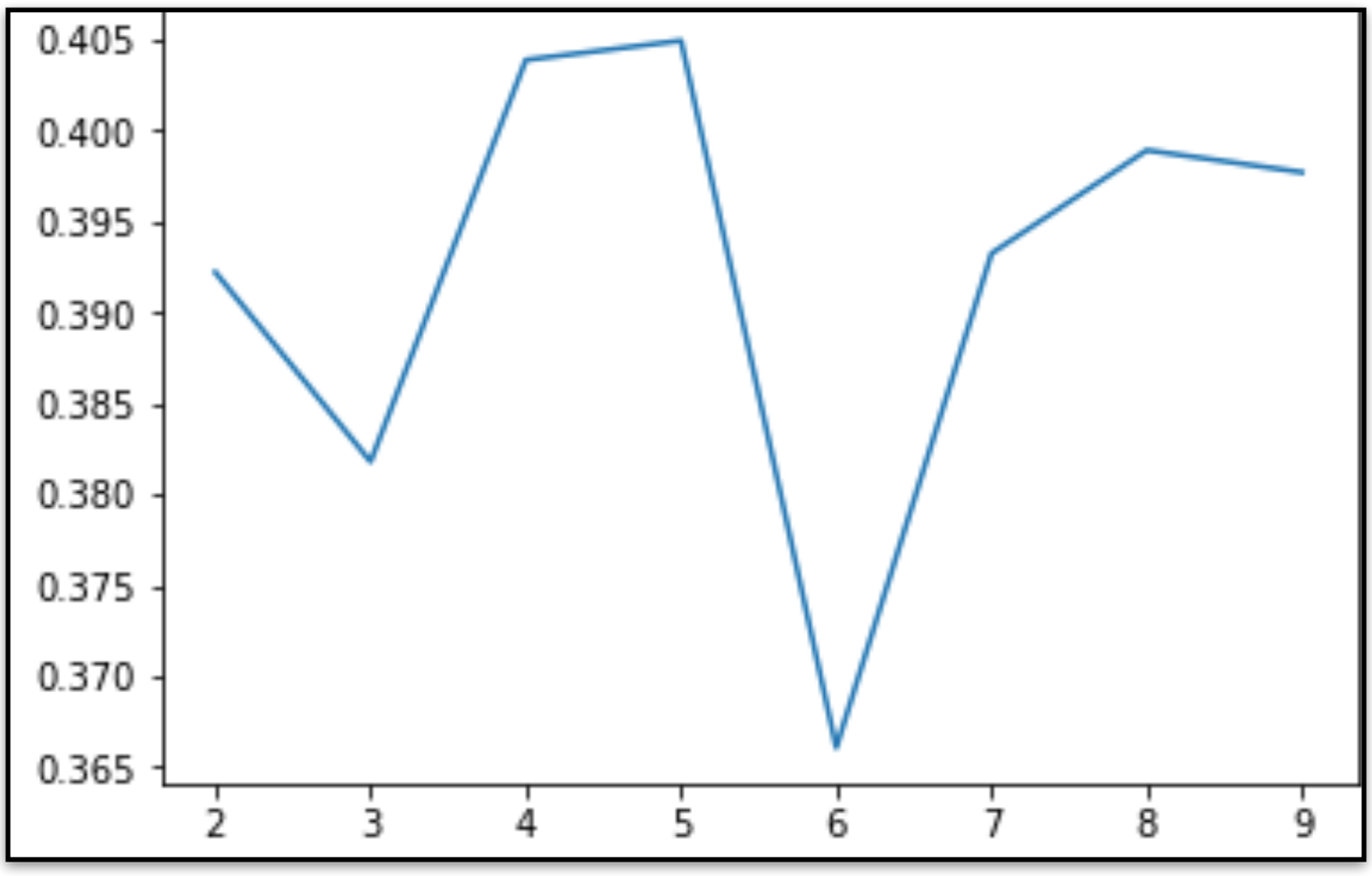
CLUSTERING

CLUSTERING OF COUNTRIES

K-Means Clustering

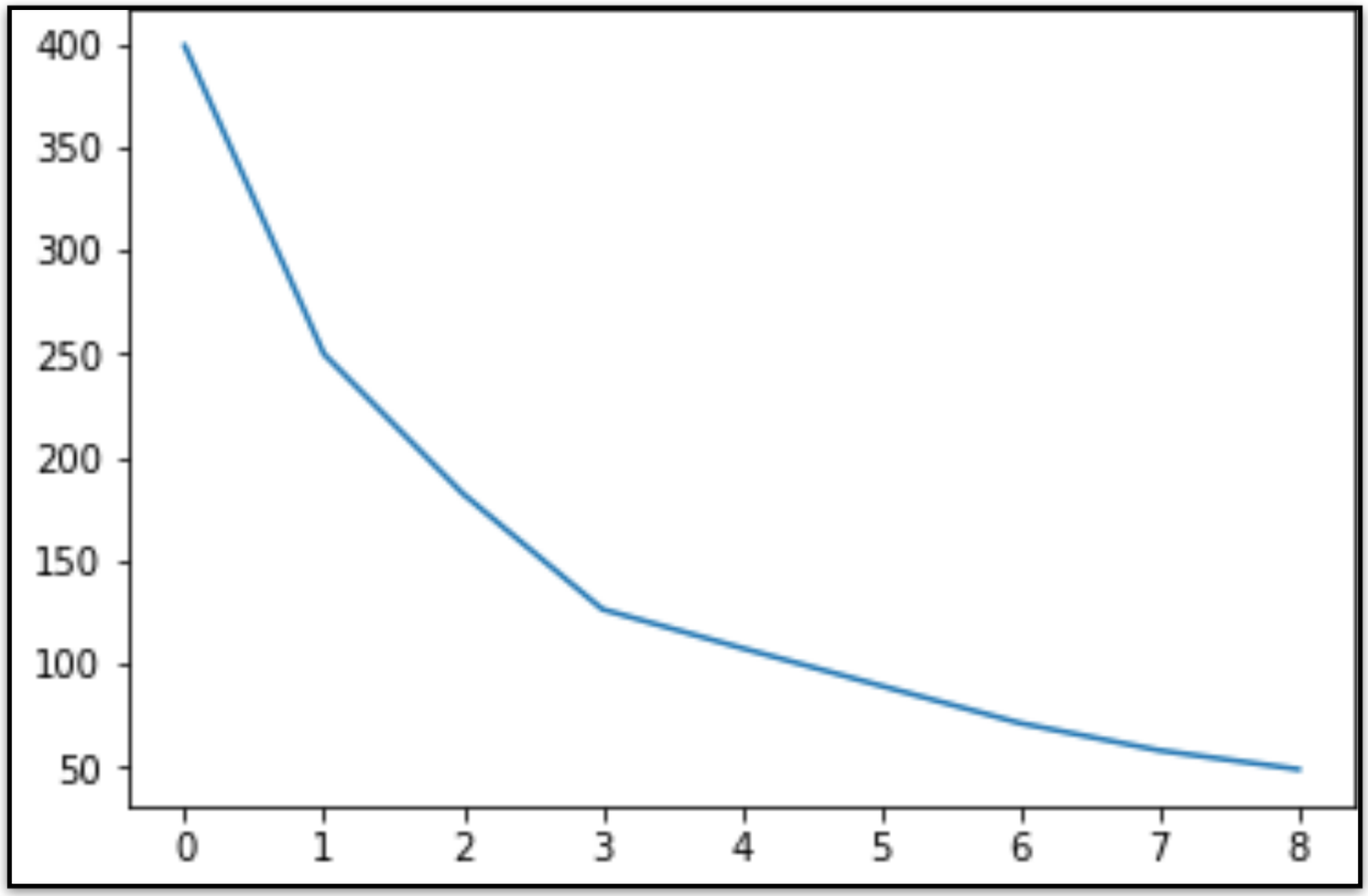
Before proceeding into clustering the data, we find out the optimal number of clusters by the following two methods:

1. Silhouette Score



Conclusion:
The silhouette score reaches a peak at around 5 clusters indicating that it might be the ideal number of clusters. (k=5)

2. Elbow Method



Conclusion:
A distinct elbow is formed at around 3-7 clusters. Making the 3rd cluster on the plot as number of clusters as 5 (K =5)

CLUSTERING OF COUNTRIES

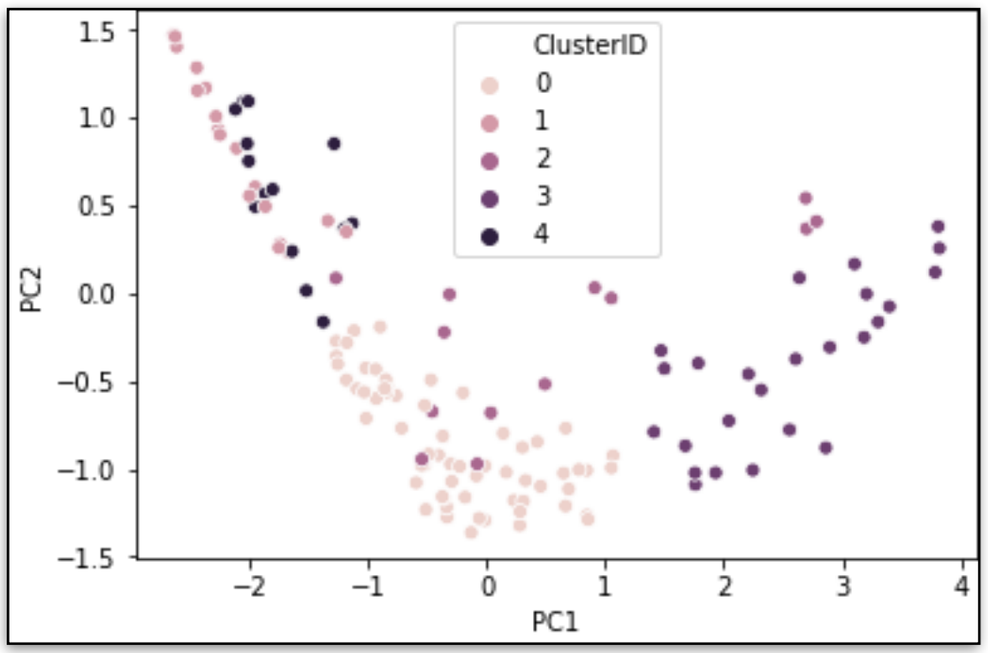
Proceeding ahead with 5 Clusters we get the following number of records in each clusters and our data head looks as follows:

Cluster ID	Value Counts
0	63
3	25
1	18
4	14
2	13

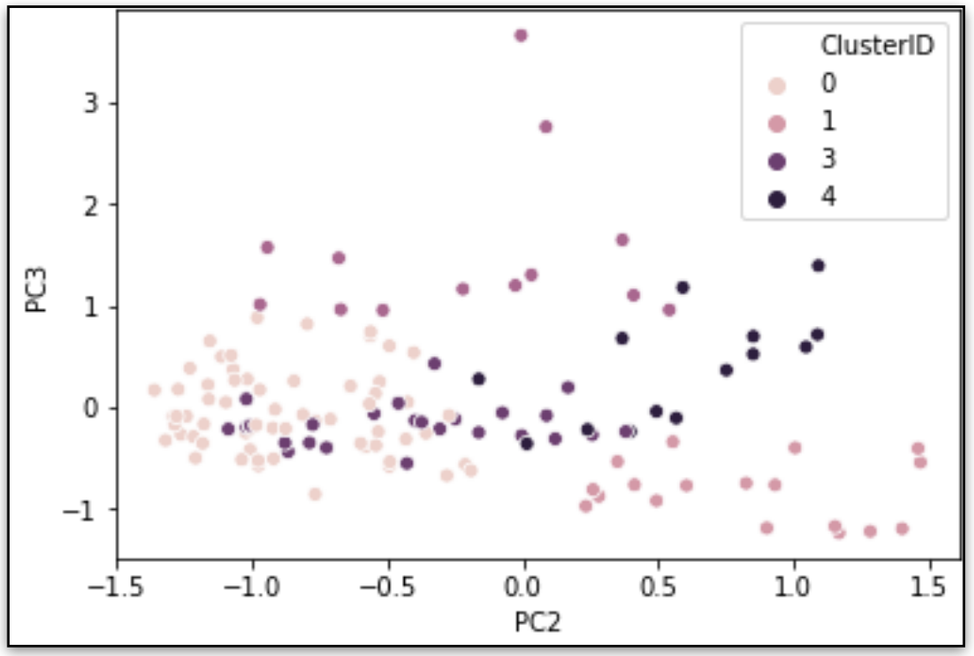
	country	PC1	PC2	PC3	ClusterID
0	Afghanistan	-2.637442	1.469038	-0.541359	1
1	Algeria	-0.457626	-0.673301	0.961867	2
2	Antigua and Barbuda	0.649849	-1.024374	-0.250103	0
3	Argentina	0.037197	-0.680889	1.466963	2
4	Armenia	-0.332692	-1.274517	0.176636	0

CLUSTERING OF COUNTRIES

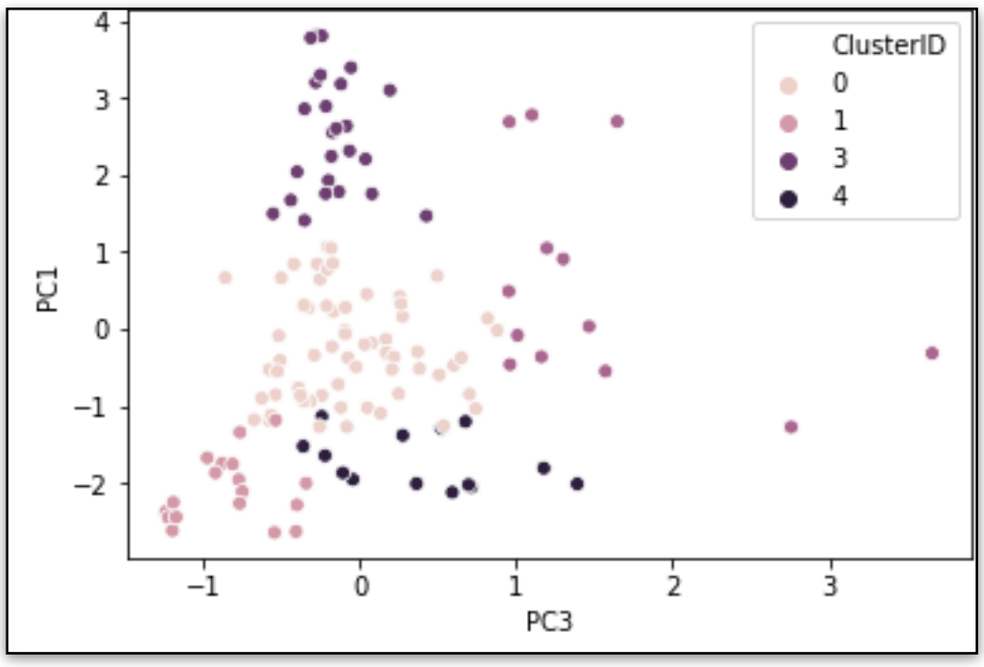
Lets have a look at the clusters formed:



PC1 vs PC2



PC2 vs PC3



PC1 vs PC3

CLUSTERING OF COUNTRIES

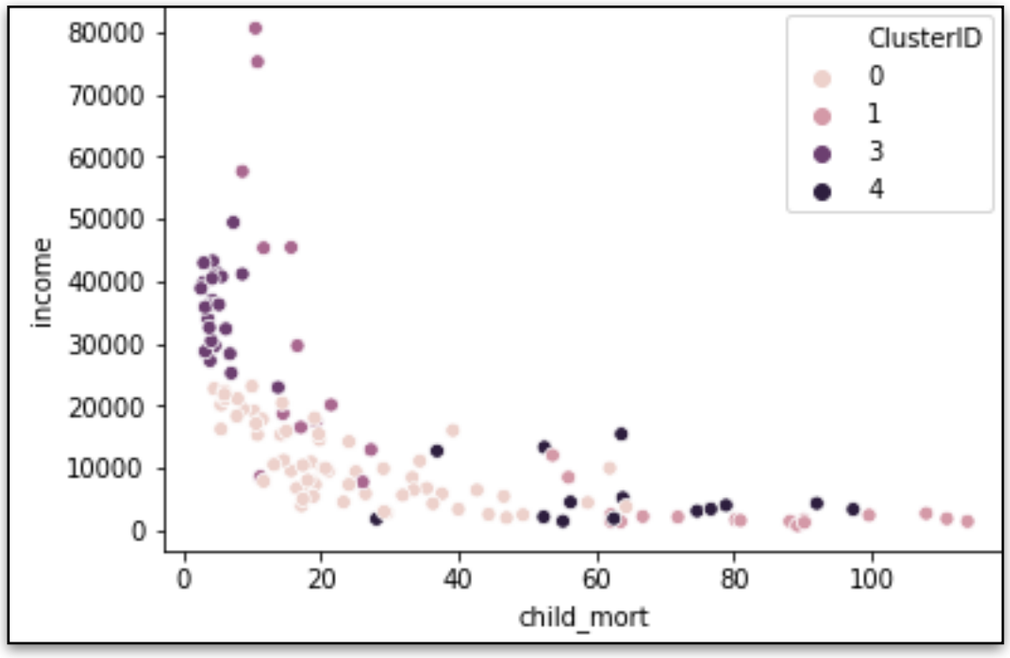
One thing we noticed is some even though some distinct clusters are being formed, some are not so good.

Now let's create the cluster means w.r.t to the various variables mentioned in the study and plot and see how they are related

	ClusterID	Child_Mortality	Exports	Imports	Health_Spending	Income	Inflation	Life_Expectancy	Total_Fertility	GDPpcapita
0	0	23.538095	2629.653437	2877.322430	376.250603	10859.682540	5.808905	72.346032	2.322857	5830.857143
1	1	80.616667	400.130500	625.700222	105.735689	2670.000000	4.488611	59.455556	4.730000	1371.333333
2	2	16.223077	9365.739231	5763.193846	599.447615	33505.384615	20.592308	74.638462	2.393846	16063.076923
3	3	4.972000	13549.368000	13383.932000	3417.322400	34832.000000	1.432920	79.972000	1.726000	34532.000000
4	4	63.671429	1070.977129	953.021714	152.232614	5445.714286	15.645000	64.642857	4.347143	2566.428571

Conclusion:

We observe that **Child mortality, Income, Inflation and GDP per capita** are good predictors for the development of a country. On cross-checking with the original Principal components that we drew, these 4 components had good scores. Hence we can say that they are a good source of information for the Clustering Process.



CLUSTERING OF COUNTRIES

Cluster Analysis

1. We observe that the best country cluster is cluster 3 based on our four important columns.

	country	child_mort	exports	imports	health	income	inflation	life_expec	total_fer	gdpp	ClusterID
5	Australia	4.8	10276.2	10847.1	4530.87	41400	1.160	82.0	1.93	51900	3
6	Austria	4.3	24059.7	22418.2	5159.00	43200	0.873	80.5	1.44	46900	3
8	Bahamas	13.8	9800.0	12236.0	2209.20	22900	-0.393	73.8	1.86	28000	3
9	Bahrain	8.6	14386.5	10536.3	1028.79	41100	7.440	76.0	2.16	20700	3
23	Canada	5.6	13793.4	14694.0	5356.20	40700	2.870	81.3	1.63	47400	3
31	Cyprus	3.6	15461.6	17710.0	1838.76	33900	2.010	79.9	1.42	30800	3
32	Czech Republic	3.4	13068.0	12454.2	1560.24	28300	-1.430	77.5	1.51	19800	3
40	Finland	3.0	17879.4	17278.8	4134.90	39800	0.351	80.0	1.87	46200	3
41	France	4.2	10880.8	11408.6	4831.40	36900	1.050	81.4	2.03	40600	3
45	Germany	4.2	17681.4	15507.8	4848.80	40400	0.758	80.1	1.39	41800	3
47	Greece	3.9	5944.9	8258.3	2770.70	28700	0.673	80.4	1.48	26900	3
53	Iceland	2.6	22374.6	18142.7	3938.60	38800	5.470	82.0	2.20	41900	3
58	Israel	4.6	10710.0	10067.4	2334.78	29600	1.770	81.4	3.03	30600	3
59	Italy	4.0	9021.6	9737.6	3411.74	36200	0.319	81.7	1.46	35800	3
61	Japan	3.2	6675.0	6052.0	4223.05	35800	-1.900	82.8	1.39	44500	3
77	Malta	6.8	32283.0	32494.0	1825.15	28300	3.830	80.3	1.36	21100	3
87	New Zealand	6.2	10211.1	9436.0	3403.70	32300	3.730	80.9	2.17	33700	3
95	Portugal	3.9	6727.5	8415.0	2475.00	27200	0.643	79.8	1.39	22500	3
103	Slovak Republic	7.0	12665.8	12914.8	1459.14	25200	0.485	75.5	1.43	16600	3
104	Slovenia	3.2	15046.2	14718.6	2201.94	28700	-0.987	79.5	1.57	23400	3
107	South Korea	4.1	10917.4	10210.2	1531.53	30400	3.160	80.1	1.23	22100	3
108	Spain	3.8	7828.5	8227.6	2928.78	32500	0.160	81.9	1.37	30700	3
113	Sweden	3.0	24070.2	21204.7	5017.23	42900	0.991	81.5	1.98	52100	3
125	United Kingdom	5.2	10969.8	11981.2	3749.96	36200	1.570	80.3	1.92	38900	3
126	United States	7.3	6001.6	7647.2	8663.60	49400	1.220	78.7	1.93	48400	3

Cluster 3

CLUSTERING OF COUNTRIES

2. We observed that cluster 1 and 4 have pretty low values of the 4 indicators that we chose. Hence these are the countries that we need to focus.

	country	child_mort	exports	imports	health	income	inflation	life_expec	total_fer	gdpp	ClusterID
0	Afghanistan	90.2	55.300	248.297	41.9174	1610	9.440	56.2	5.82	553	1
14	Benin	111.0	180.404	281.976	31.0780	1820	0.885	61.8	5.36	758	1
22	Cameroon	108.0	290.820	353.700	67.2030	2660	1.910	57.3	5.11	1310	1
28	Comoros	88.2	126.885	397.573	34.6819	1410	3.870	65.9	4.75	769	1
43	Gambia	80.3	133.756	239.974	31.9778	1660	4.300	65.5	5.71	562	1
50	Guinea-Bissau	114.0	81.503	192.544	46.4950	1390	2.970	55.6	5.05	547	1
64	Kenya	62.2	200.169	324.912	45.9325	2480	2.090	62.8	4.37	967	1
65	Kiribati	62.7	198.170	1190.510	168.3700	1730	1.520	60.7	3.84	1490	1
70	Lesotho	99.7	460.980	1181.700	129.8700	2380	4.150	46.5	3.30	1170	1
71	Liberia	89.3	62.457	302.802	38.5860	700	5.470	60.8	5.02	327	1
74	Madagascar	62.2	103.250	177.590	15.5701	1390	8.790	60.8	4.60	413	1
85	Namibia	56.0	2480.820	3150.330	351.8820	8460	3.560	58.6	3.60	5190	1
98	Rwanda	63.6	67.560	168.900	59.1150	1350	2.610	64.6	4.51	563	1
101	Senegal	66.8	249.000	403.000	56.6000	2180	1.850	64.0	5.06	1000	1
106	South Africa	53.7	2082.080	1994.720	650.8320	12000	6.350	54.3	2.59	7280	1
115	Tanzania	71.9	131.274	204.282	42.1902	2090	9.250	59.3	5.43	702	1
117	Togo	90.3	196.176	279.624	37.3320	1210	1.180	58.7	4.87	488	1
122	Uganda	81.0	101.745	170.170	53.6095	1540	10.600	56.8	6.15	595	1

Cluster 1

CLUSTERING OF COUNTRIES

	country	child_mort	exports	imports	health	income	inflation	life_expec	total_fer	gdpp	ClusterID
17	Botswana	52.5	2768.6000	3257.550	527.0500	13300	8.92	57.1	2.88	6350	4
29	Congo, Rep.	63.9	2331.7400	1498.780	67.4040	5190	20.70	60.4	4.95	2740	4
37	Eritrea	55.2	23.0878	112.306	12.8212	1420	11.60	61.7	4.61	482	4
42	Gabon	63.7	5048.7500	1653.750	306.2500	15400	16.60	62.9	4.08	8750	4
46	Ghana	74.7	386.4500	601.290	68.3820	3060	16.60	62.2	4.27	1310	4
57	Iraq	36.9	1773.0000	1534.500	378.4500	12700	16.60	67.2	4.56	4500	4
68	Lao	78.9	403.5600	562.020	50.9580	3980	9.20	63.8	3.15	1140	4
78	Mauritania	97.4	608.4000	734.400	52.9200	3320	18.90	68.2	4.98	1200	4
89	Pakistan	92.1	140.4000	201.760	22.8800	4280	10.90	65.3	3.85	1040	4
105	Solomon Islands	28.1	635.9700	1047.480	110.2950	1780	6.81	61.7	4.24	1290	4
111	Sudan	76.7	291.5600	254.560	93.5360	3370	19.60	66.3	4.88	1480	4
114	Tajikistan	52.4	109.9620	432.468	44.1324	2110	12.50	69.6	3.51	738	4
116	Timor-Leste	62.6	79.2000	1000.800	328.3200	1850	26.50	71.1	6.23	3600	4
132	Yemen	56.3	393.0000	450.640	67.8580	4480	23.60	67.5	4.67	1310	4

Cluster 4

Hierarchical Clustering

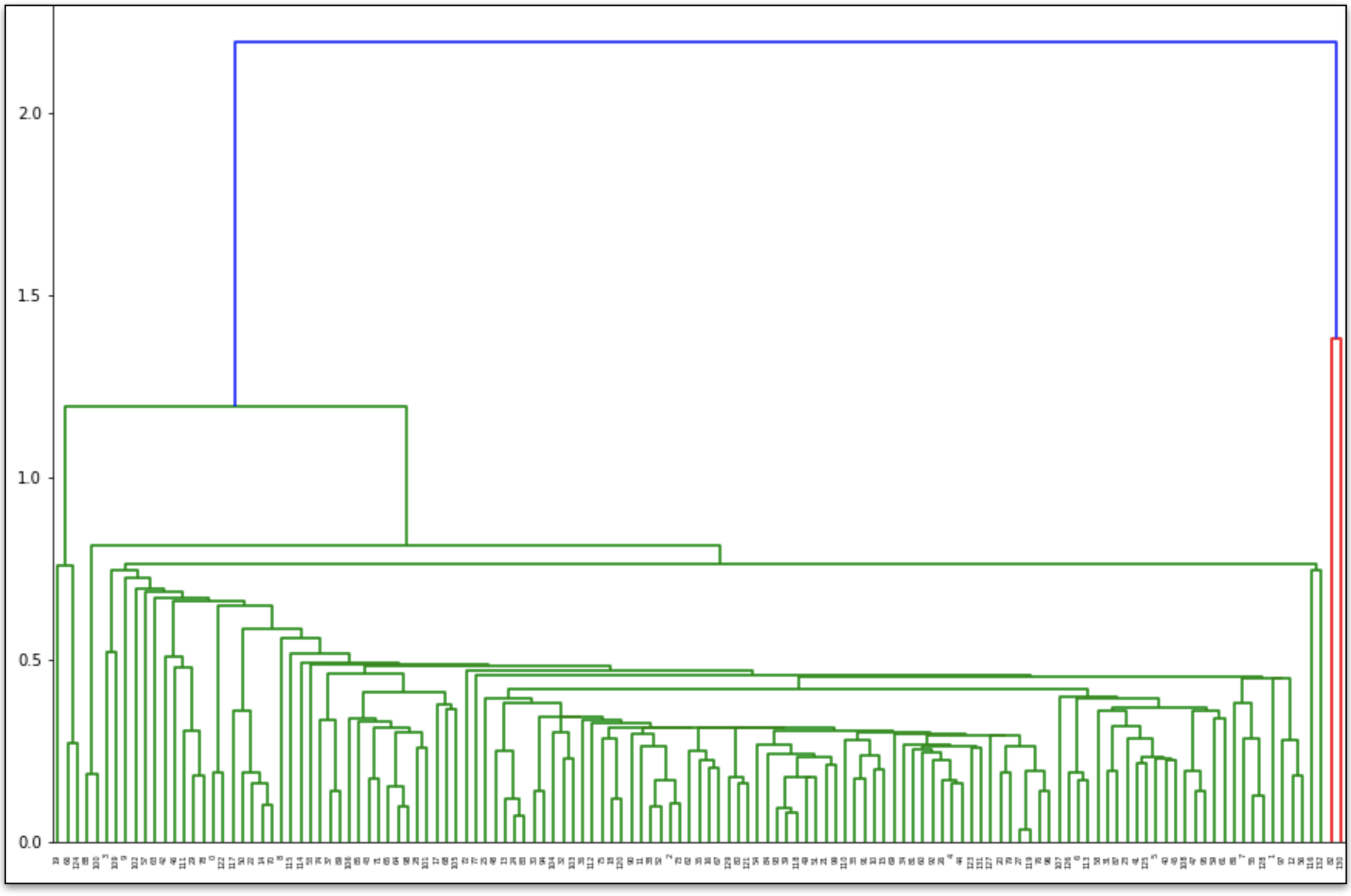
Hierarchical Clustering has one advantage over K-means Clustering which is that we don't have to select the initial number of clusters before performing clustering.

It has a different concept of linkage through which it performs the clustering operations. There are two types of Linkage:

1. Single Linkage
2. Complete Linkage

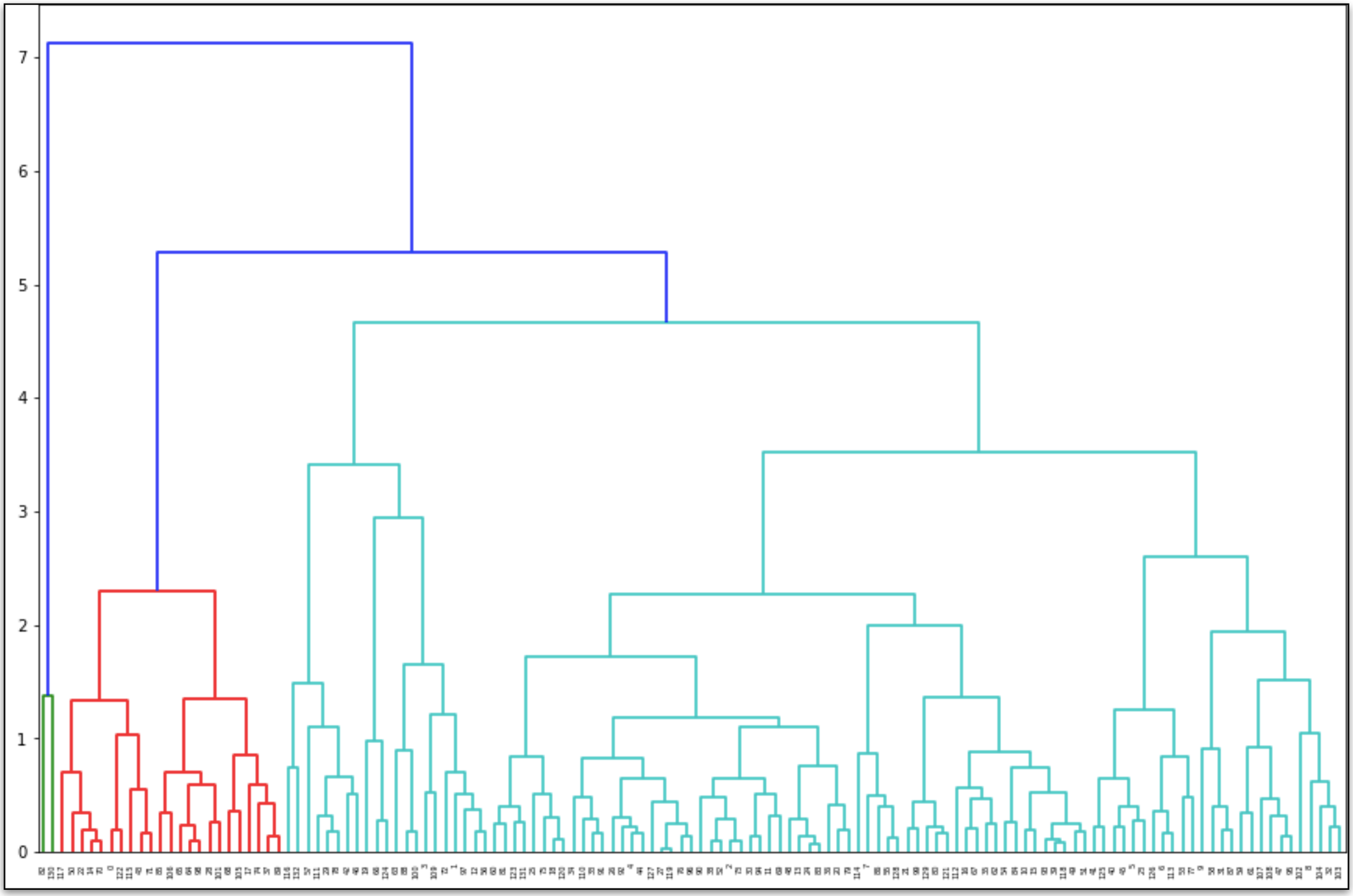
Lets try both the methods on our country data and see if the results are good enough.

CLUSTERING OF COUNTRIES



Single Linkage do not give clear cluster formation so we have to try complete linkage in the next step.

CLUSTERING OF COUNTRIES



Now we see some good amount of clusters getting formed.

CLUSTERING OF COUNTRIES

Now if we cut the tree at 5 clusters and look at our data head after assigning the cluster ids. The results are as such:

Cluster ID	Value Counts
2	61
3	26
0	23
1	21
4	2

	country	PC1	PC2	PC3	ClusterID
0	Afghanistan	-2.637442	1.469038	-0.541359	0
1	Algeria	-0.457626	-0.673301	0.961867	1
2	Antigua and Barbuda	0.649849	-1.024374	-0.250103	2
3	Argentina	0.037197	-0.680889	1.466963	1
4	Armenia	-0.332692	-1.274517	0.176636	2

Principal Component Data Head

	country	child_mort	exports	imports	health	income	inflation	life_expec	total_fer	gdpp	ClusterID
0	Afghanistan	90.2	55.30	248.297	41.9174	1610	9.44	56.2	5.82	553	0
1	Algeria	27.3	1712.64	1400.440	185.9820	12900	16.10	76.5	2.89	4460	1
2	Antigua and Barbuda	10.3	5551.00	7185.800	735.6600	19100	1.44	76.8	2.13	12200	2
3	Argentina	14.5	1946.70	1648.000	834.3000	18700	20.90	75.8	2.37	10300	1
4	Armenia	18.1	669.76	1458.660	141.6800	6700	7.77	73.3	1.69	3220	2

Original Dataset Data Head

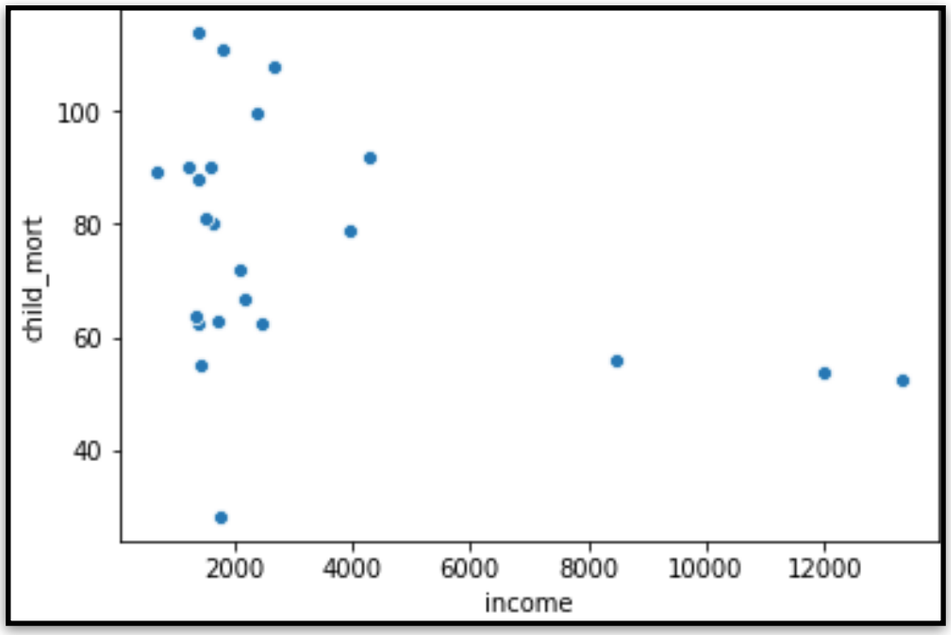
CLUSTERING OF COUNTRIES

1. We observe that the best country cluster is cluster 3 based on our four important columns.

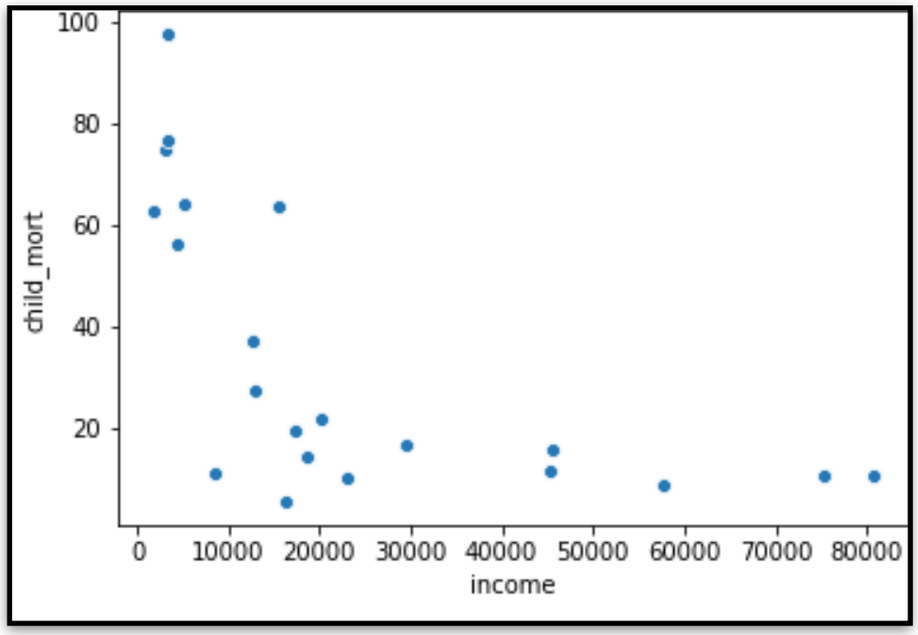
	country	child_mort	exports	imports	health	income	inflation	life_expec	total_fer	gdpp	ClusterID
5	Australia	4.8	10276.2	10847.1	4530.87	41400	1.160	82.0	1.93	51900	3
6	Austria	4.3	24059.7	22418.2	5159.00	43200	0.873	80.5	1.44	46900	3
8	Bahamas	13.8	9800.0	12236.0	2209.20	22900	-0.393	73.8	1.86	28000	3
9	Bahrain	8.6	14386.5	10536.3	1028.79	41100	7.440	76.0	2.16	20700	3
23	Canada	5.6	13793.4	14694.0	5356.20	40700	2.870	81.3	1.63	47400	3
31	Cyprus	3.6	15461.6	17710.0	1838.76	33900	2.010	79.9	1.42	30800	3
32	Czech Republic	3.4	13068.0	12454.2	1560.24	28300	-1.430	77.5	1.51	19800	3
40	Finland	3.0	17879.4	17278.8	4134.90	39800	0.351	80.0	1.87	46200	3
41	France	4.2	10880.8	11408.6	4831.40	36900	1.050	81.4	2.03	40600	3
45	Germany	4.2	17681.4	15507.8	4848.80	40400	0.758	80.1	1.39	41800	3
47	Greece	3.9	5944.9	8258.3	2770.70	28700	0.673	80.4	1.48	26900	3
53	Iceland	2.6	22374.6	18142.7	3938.60	38800	5.470	82.0	2.20	41900	3
58	Israel	4.6	10710.0	10067.4	2334.78	29600	1.770	81.4	3.03	30600	3
59	Italy	4.0	9021.6	9737.6	3411.74	36200	0.319	81.7	1.46	35800	3
61	Japan	3.2	6675.0	6052.0	4223.05	35800	-1.900	82.8	1.39	44500	3
77	Malta	6.8	32283.0	32494.0	1825.15	28300	3.830	80.3	1.36	21100	3
87	New Zealand	6.2	10211.1	9436.0	3403.70	32300	3.730	80.9	2.17	33700	3
95	Portugal	3.9	6727.5	8415.0	2475.00	27200	0.643	79.8	1.39	22500	3
102	Seychelles	14.4	10130.4	11664.0	367.20	20400	-4.210	73.4	2.17	10800	3
103	Slovak Republic	7.0	12665.8	12914.8	1459.14	25200	0.485	75.5	1.43	16600	3
104	Slovenia	3.2	15046.2	14718.6	2201.94	28700	-0.987	79.5	1.57	23400	3
107	South Korea	4.1	10917.4	10210.2	1531.53	30400	3.160	80.1	1.23	22100	3
108	Spain	3.8	7828.5	8227.6	2928.78	32500	0.160	81.9	1.37	30700	3
113	Sweden	3.0	24070.2	21204.7	5017.23	42900	0.991	81.5	1.98	52100	3
125	United Kingdom	5.2	10969.8	11981.2	3749.96	36200	1.570	80.3	1.92	38900	3

CLUSTERING OF COUNTRIES

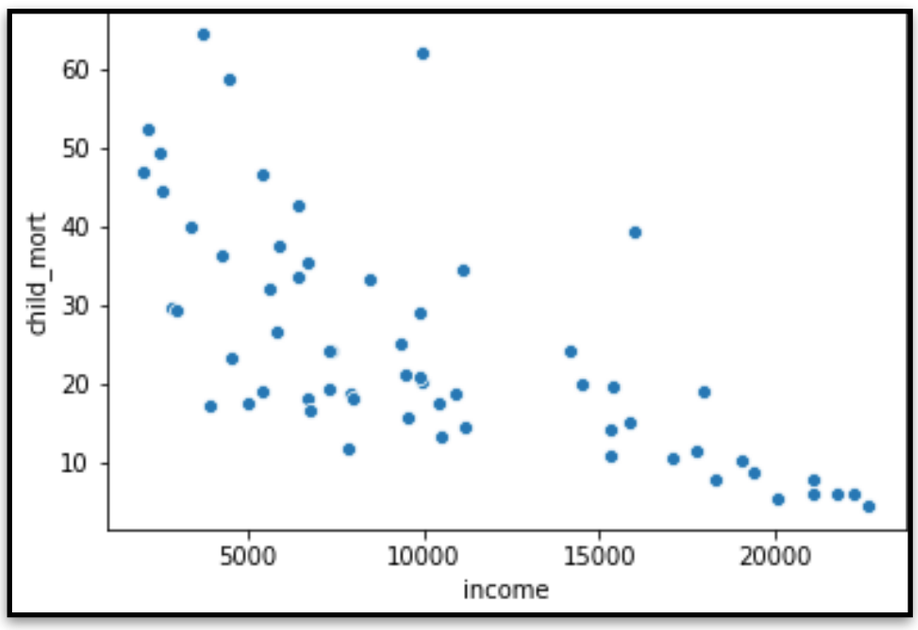
2. We observe that the clustering on the other clusters are not so prominent as it was in K-means.



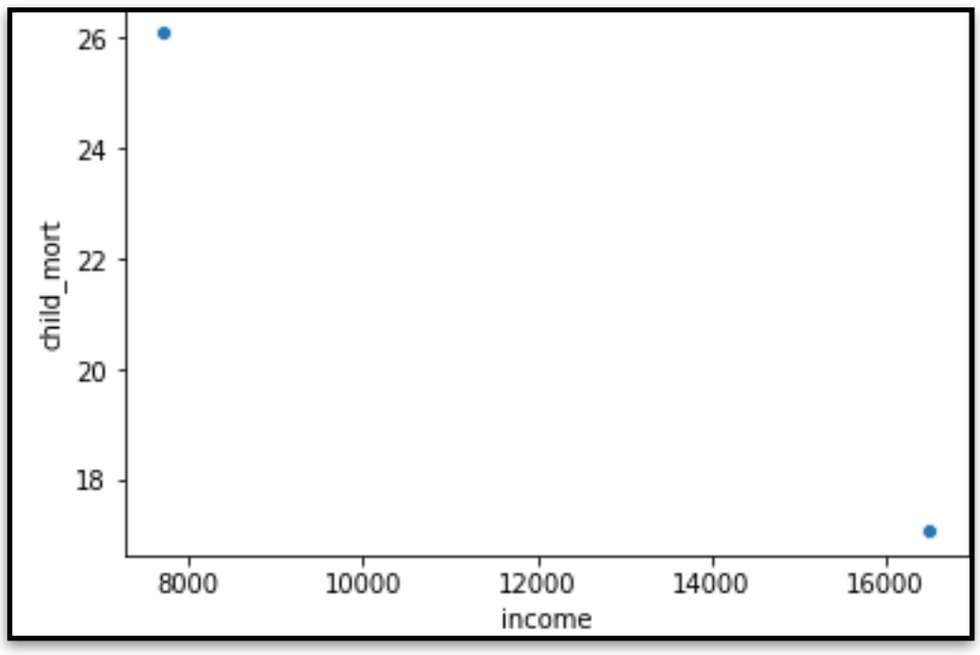
Cluster 0



Cluster 1



Cluster 2



Cluster 4

Conclusion

We use the clusters formed during K-means clustering to find the countries that we require since Hierarchical clustering is not showing proper clusters here. For K-means part, we got Cluster 1 and 4 might be the ones which has a proper need of aid. Cluster 2 countries might also improve with aid.

In this solution K-means and Hierarchical don't produce identical insights.This would depend on the way the principal components and the final number of clusters are chosen. It would be perfectly fine if both provide identical insights in any other case

For outliers that I removed before clustering took place, I can take any approach to include them in the final list of countries that I'd focus on. Either reassign them to the clusters that were formed and see if Cluster 2 and Cluster 4 have any more countries or use one variable from some of the main indicators to bin the entire 167 countries. Like if I take GDPP, keep the bin limits as 0-1700, 1700-3200, 3200-6000, 6000-13000 and >13000. The bin limits are decided on the basis of the approximate gdpp means that we got for the 5 clusters. Categorise all the countries from the original dataset within these limits and then take all the countries less than 1700 as the cutoff.Similarly, proceed for the 2nd variable.Any other logically thought out approach also works.

CLUSTERING OF COUNTRIES

FINAL LIST OF COUNTRIES TO FOCUS ON

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
0	Afghanistan	90.2	55.3000	41.9174	248.297	1610	9.440	56.2	5.82	553
17	Benin	111.0	180.4040	31.0780	281.976	1820	0.885	61.8	5.36	758
25	Burkina Faso	116.0	110.4000	38.7550	170.200	1430	6.810	57.9	5.87	575
26	Burundi	93.6	20.6052	26.7960	90.552	764	12.300	57.7	6.26	231
28	Cameroon	108.0	290.8200	67.2030	353.700	2660	1.910	57.3	5.11	1310
31	Central African Republic	149.0	52.6280	17.7508	118.190	888	2.010	47.5	5.21	446
32	Chad	150.0	330.0960	40.6341	390.195	1930	6.390	56.5	6.59	897
36	Comoros	88.2	126.8850	34.6819	397.573	1410	3.870	65.9	4.75	769
37	Congo, Dem. Rep.	116.0	137.2740	26.4194	165.664	609	20.800	57.5	6.54	334
40	Cote d'Ivoire	111.0	617.3200	64.6600	528.260	2690	5.390	56.3	5.27	1220
56	Gambia	80.3	133.7560	31.9778	239.974	1660	4.300	65.5	5.71	562
63	Guinea	109.0	196.3440	31.9464	279.936	1190	16.100	58.0	5.34	648
64	Guinea-Bissau	114.0	81.5030	46.4950	192.544	1390	2.970	55.6	5.05	547
66	Haiti	208.0	101.2860	45.7442	428.314	1500	5.450	32.1	3.33	662
87	Lesotho	99.7	460.9800	129.8700	1181.700	2380	4.150	46.5	3.30	1170
88	Liberia	89.3	62.4570	38.5860	302.802	700	5.470	60.8	5.02	327
94	Malawi	90.5	104.6520	30.2481	160.191	1030	12.100	53.1	5.31	459
97	Mali	137.0	161.4240	35.2584	248.508	1870	4.370	59.5	6.55	708
106	Mozambique	101.0	131.9850	21.8299	193.578	918	7.640	54.5	5.56	419
112	Niger	123.0	77.2560	17.9568	170.868	814	2.550	58.8	7.49	348
132	Sierra Leone	160.0	67.0320	52.2690	137.655	1220	17.200	55.0	5.20	399
150	Togo	90.3	196.1760	37.3320	279.624	1210	1.180	58.7	4.87	488
155	Uganda	81.0	101.7450	53.6095	170.170	1540	10.600	56.8	6.15	595

THANK YOU