Clustering

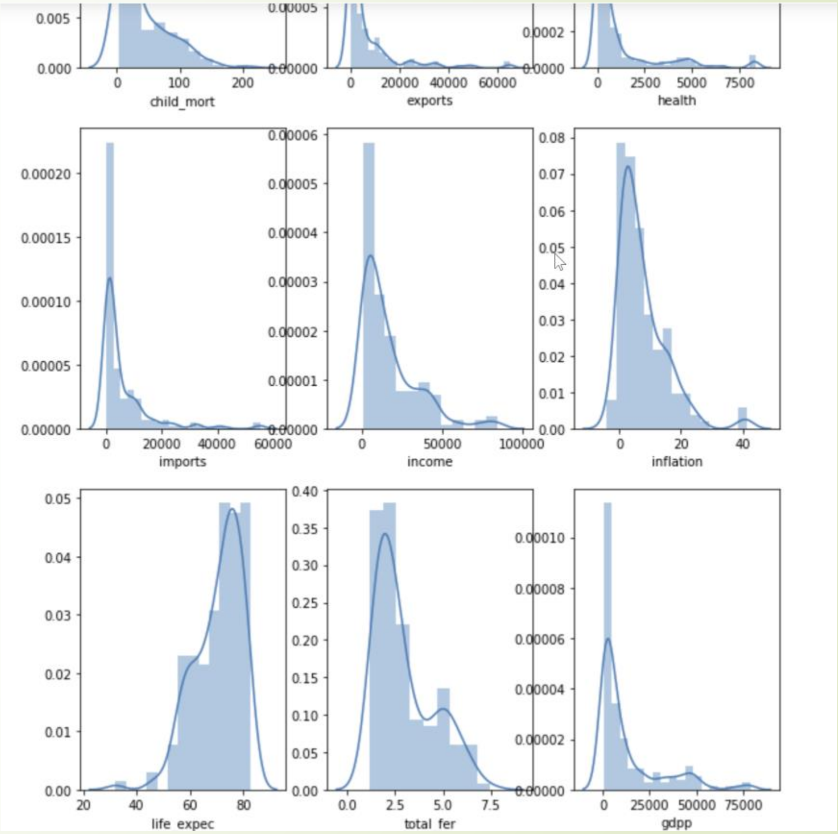
To identify top 5 countries in need of aid.

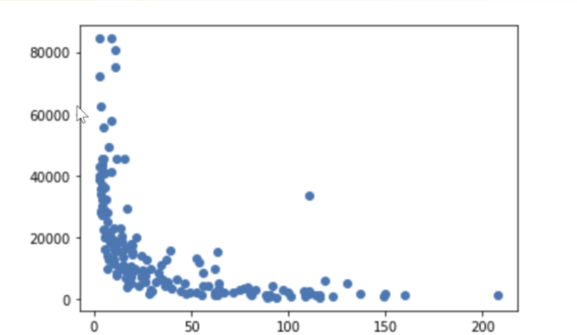
Suby Oommen

Problem statement: HELP International NGO needs a list of top 5 countries that are in the direst need of aid. To arrive at this list, we need to analyse and categorise the countries using socio-economic and health factors that determine the overall development of the country. Then list the top 5 countries for the CEO to focus on.

Important EDA steps:

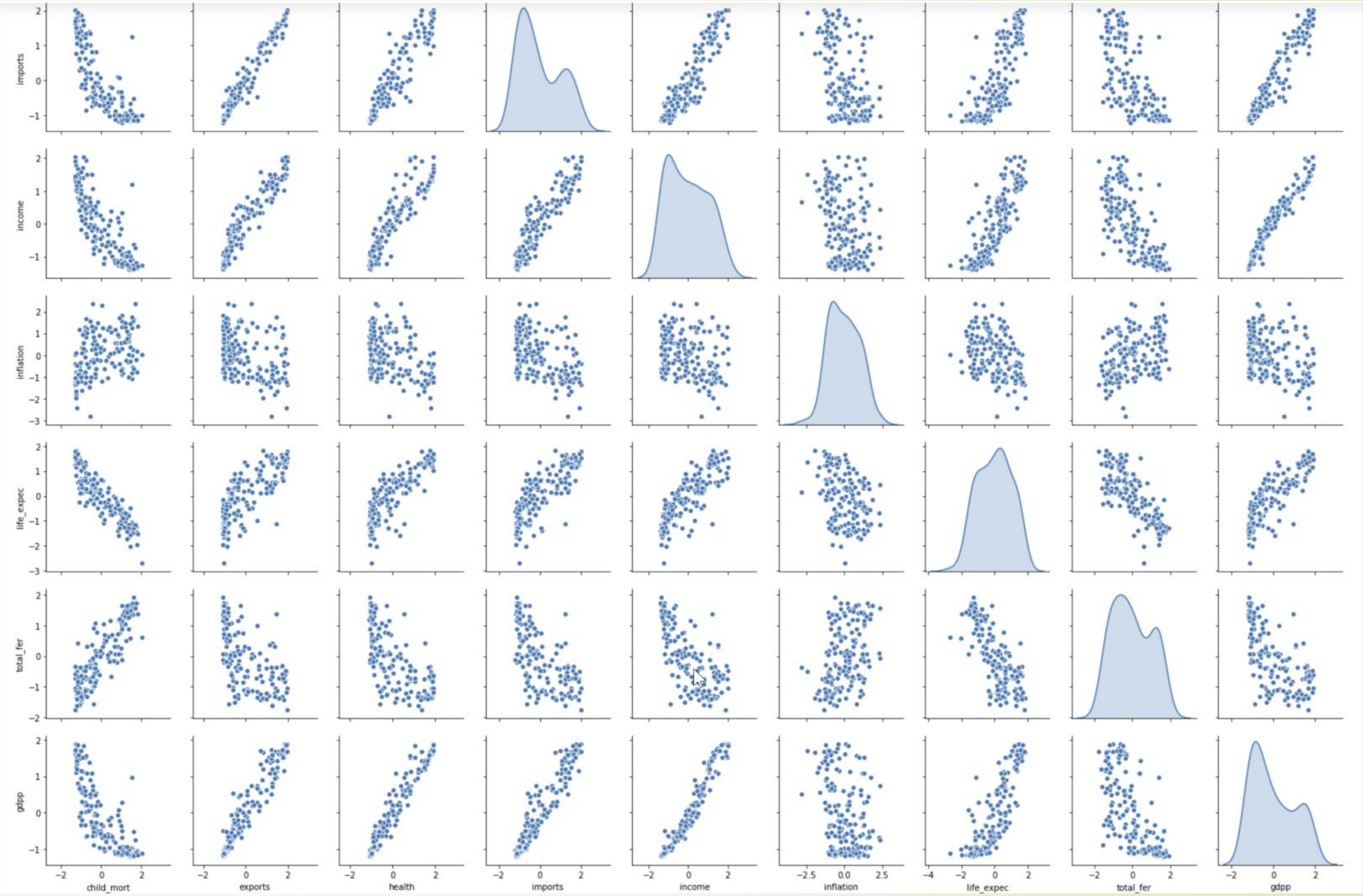
- Convert exports, imports & health features to absolute values, as they were in percentage of GDP
- Cap outliers to .99 quantile for gdpp, income, exports etc as they focus on rich countries. We should not cap 'child moratlity' as it is a relevant feature to identify needy countries.
- Visualizations assist in analysis e.g. distplot, we can observe multiple peaks in the distribution.
- Scatterplot e.g. Higher income, lower child mortality





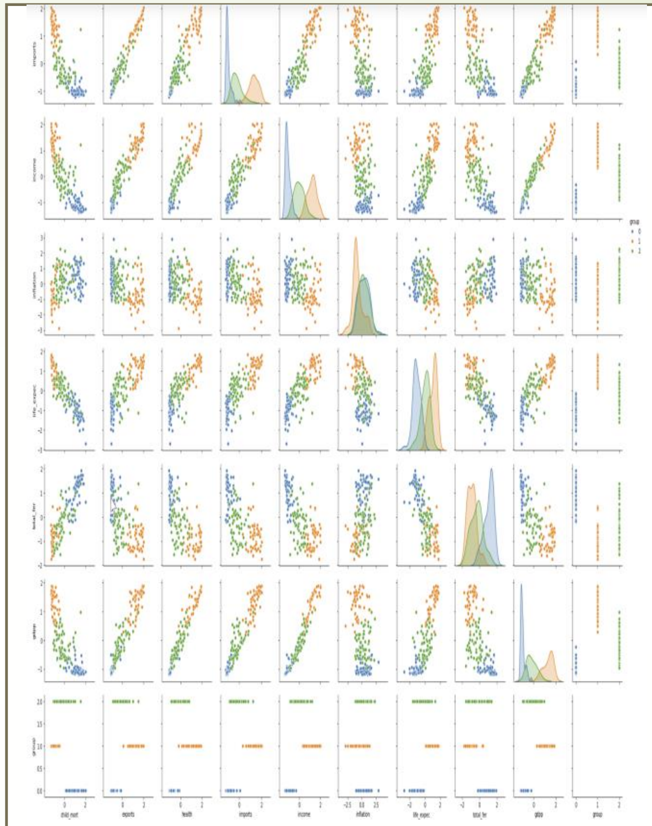
Pre-Processing

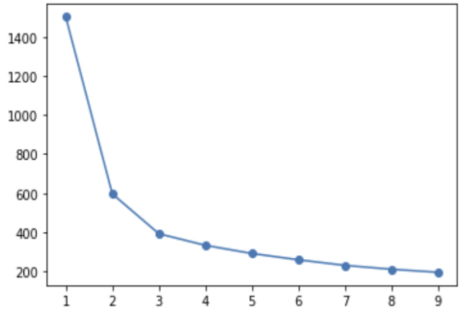
- Perform scaling to ensure no feature overshadows the other.
- PowerTransformer to attain Normal Disrtibution
- Result from sns.pairplot is shared

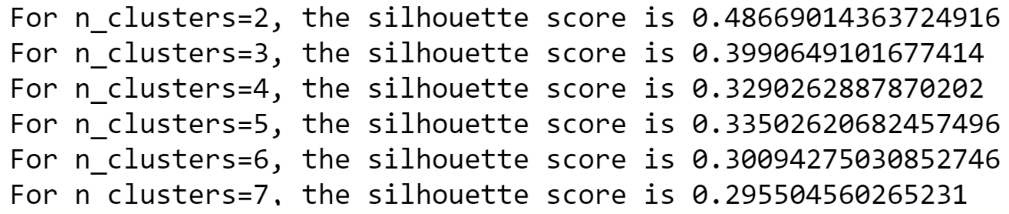


K Means

- Optimal cluster identified is **3** using the two methods below:
- **Elbow/SSD method:** The graph on the right indicates that the rate of drop after 3 is not significant, hence 3 is the optimal cluster choice.
- The **silhouette coefficient** for cluster =3 is 0.399
- Three distinct clusters are visible in the pairplot graph.



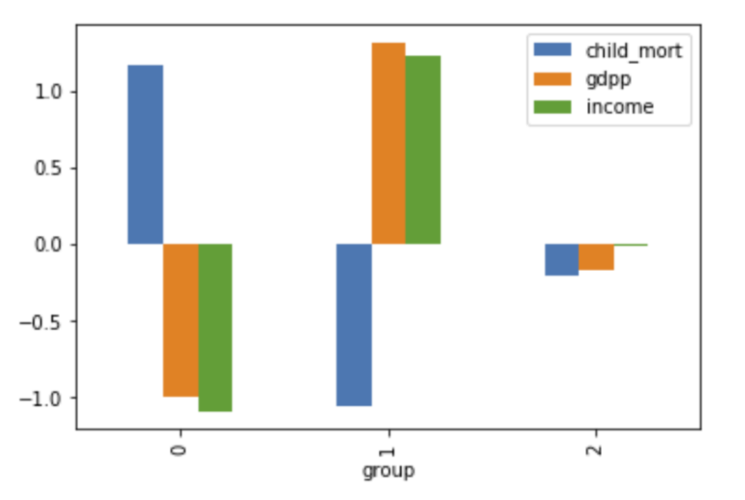


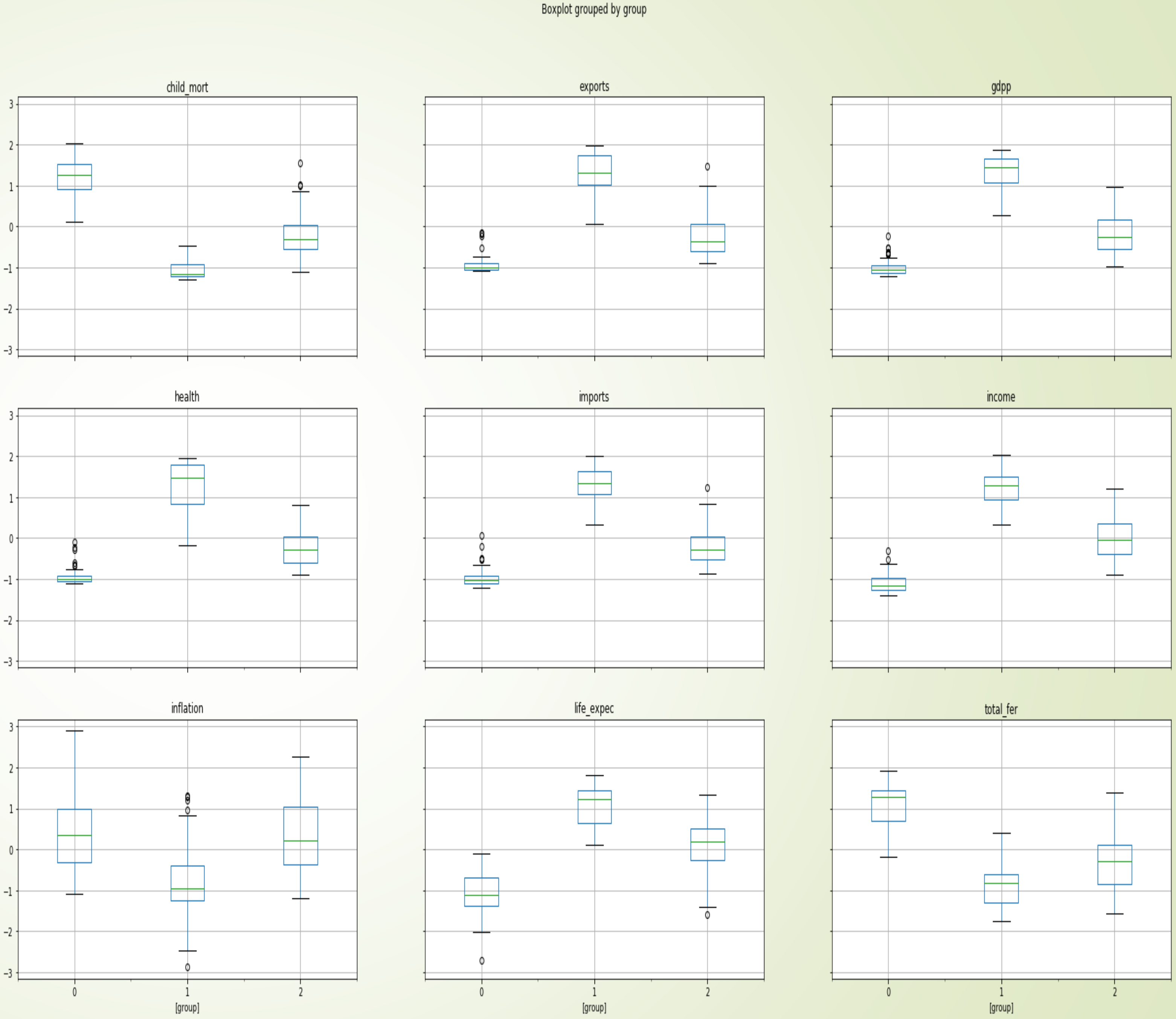


```
For n_clusters=2, the silhouette score is 0.48669014363724916
For n_clusters=3, the silhouette score is 0.3990649101677414
For n_clusters=4, the silhouette score is 0.3290262887870202
For n_clusters=5, the silhouette score is 0.33502620682457496
For n_clusters=6, the silhouette score is 0.30094275030852746
For n_clusters=7, the silhouette score is 0.295504560265231
```

Clusters:

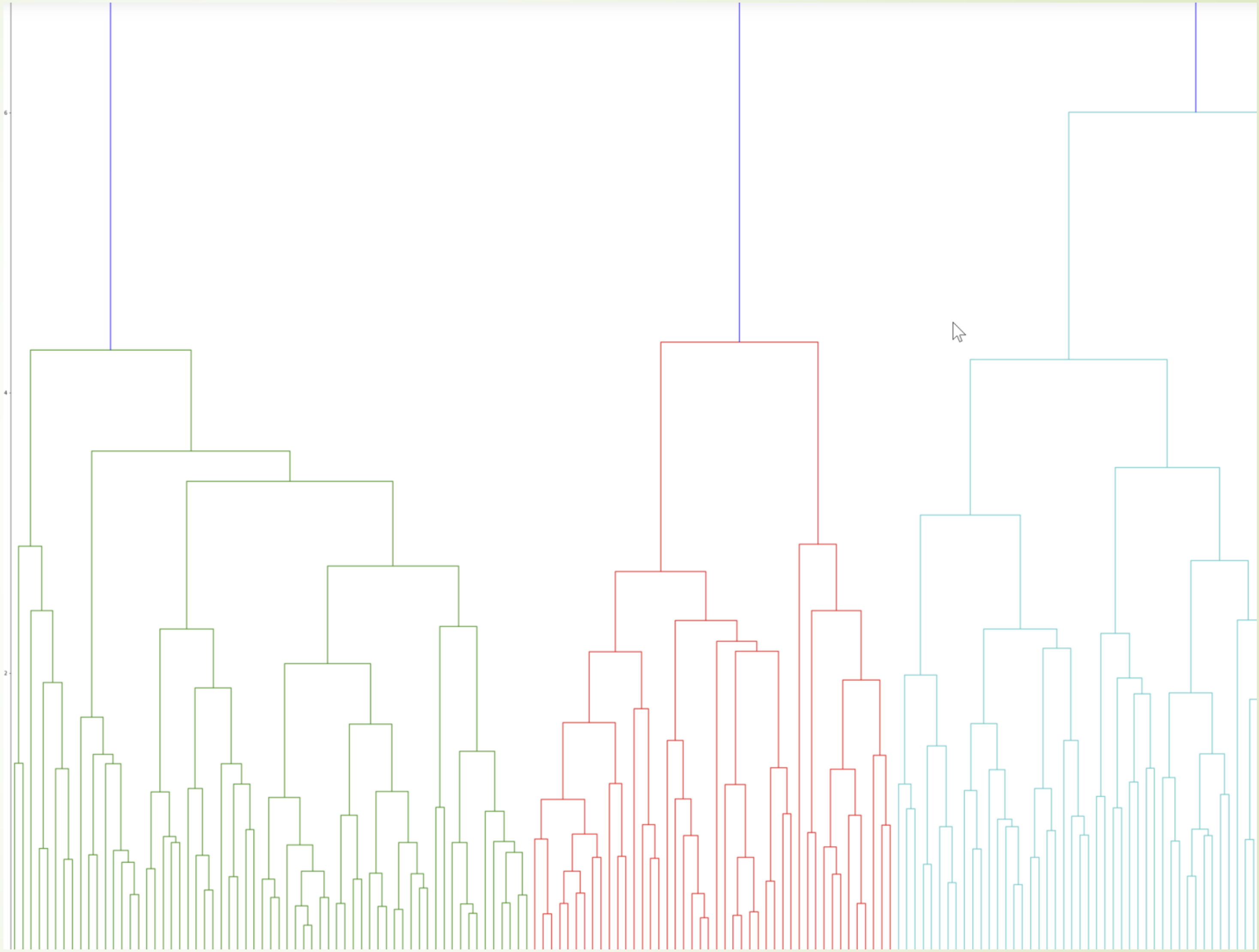
- **Group 0** - high child mortality, low gdpp, low income, low life_expec, high total_fer, low import and export
- **Group 1** - low child mortality, high gdpp, high income, low fertility, high life_expec and high import, export and health
- **Group 2** - This has values intermediate between Group 0 and Group 1



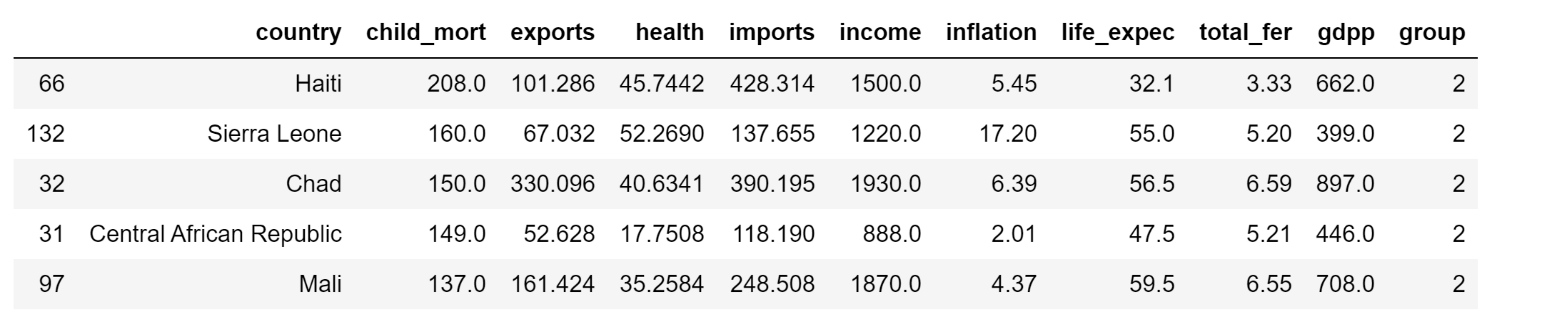


Hierarchial clustering

- Performed **Single** and **Complete** Linkages
- We can observe **3** distinct clusters in **dendogram** on Complete Linkage



Top 5 countries identified using K-Means and Hierarchical clustering



	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	group
66	Haiti	208.0	101.286	45.7442	428.314	1500.0	5.45	32.1	3.33	662.0	2
132	Sierra Leone	160.0	67.032	52.2690	137.655	1220.0	17.20	55.0	5.20	399.0	2
32	Chad	150.0	330.096	40.6341	390.195	1930.0	6.39	56.5	6.59	897.0	2
31	Central African Republic	149.0	52.628	17.7508	118.190	888.0	2.01	47.5	5.21	446.0	2
97	Mali	137.0	161.424	35.2584	248.508	1870.0	4.37	59.5	6.55	708.0	2