CLUSTERING ASSIGNMENT

Abhay Saxena

Problem Statement

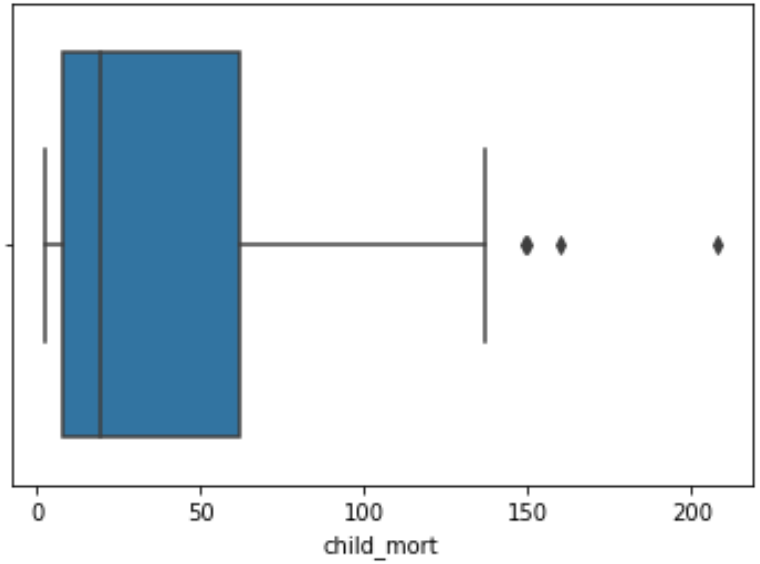
- Clustering of countries, is to categorise the countries using some socio-economic and health factors that determine the overall development of the country. Then we need to suggest the countries which the CEO needs to focus on the most and are in direst need for aid.
- This includes:
  - EDA, Outlier Analysis, Scaling, and Hopkins Test.
  - K-means and Hierarchical clustering(both single and complete linkage)
  - Analyse the clusters by comparing how these three variables - [gdpp, child_mort and income]
  - To perform visualisations on the clusters that have been formed

Approach

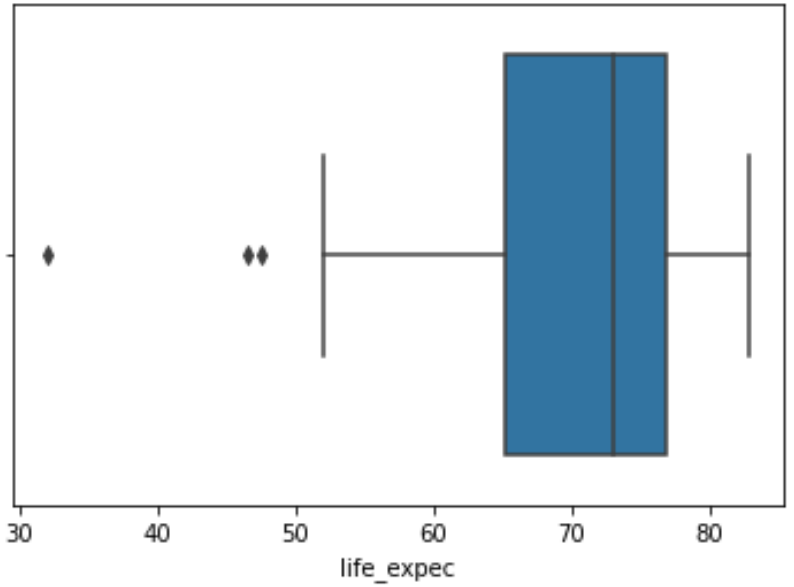
1. EDA
    1. Univariate
    2. Bivariate
    3. Outlier treatment
2. Scaling
3. Hopkins Test
4. K-means Clustering
    1. SSD
    2. Silhouette
    3. Final Model
    4. Clustering Profile
5. Hierarchical Clustering
    1. Single Linkage
    2. Complete Linkage
    3. Cluster Cutting
    4. Cluster profiling
6. Ranking list of Countries that are in dire need of Aid.

Outliers Treatment:

For columns such as child_mort, inflation, total_fer, have not done anything to the upper range outliers but have dealt with the lower range outlier(capping).
But for rest of the columns, we have not done anything for the lower range outliers but dealt with the upper range outliers(capping)



Countries with Child_mort>140 and life_expec<50 years




```
In [211]: country_df[country_df.life_expec<50].sort_values(by='health',ascending= True).head(10)
```

Out[211]:

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
31	Central African Republic	149.0	52.63	17.75	118.19	888	2.01	47.5	5.21	446
66	Haiti	208.0	101.29	45.74	428.31	1500	5.45	32.1	3.33	662
87	Lesotho	99.7	460.98	129.87	1181.70	2380	4.15	46.5	3.30	1170

Countries with Life_expec < 50 years and least amount on health spending.

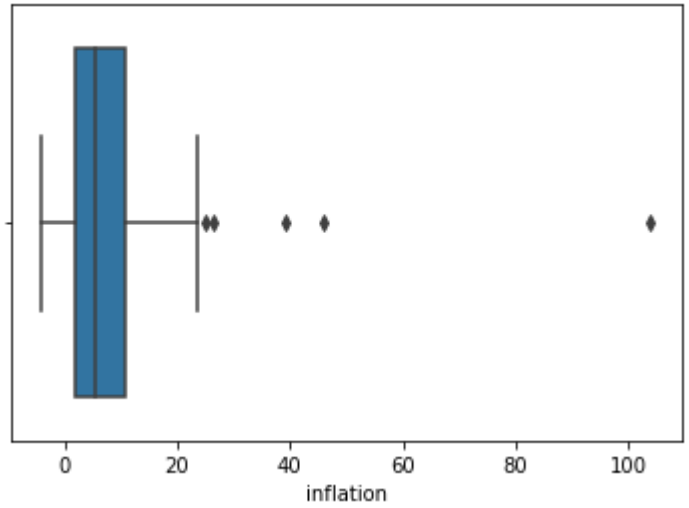
EDA

Countries Where Export < Imports & income is also low.

```
In [224]: # Countries where Export < Import and income is also low.
country_df[country_df.exports<country_df.imports].sort_values(by='income',ascending= True).head()
```

Out[224]:

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
37	Congo, Dem. Rep.	116.0	137.27	26.42	165.66	609	20.80	57.5	6.54	334
88	Liberia	89.3	62.46	38.59	302.80	700	5.47	60.8	5.02	327
26	Burundi	93.6	20.61	26.80	90.55	764	12.30	57.7	6.26	231
112	Niger	123.0	77.26	17.96	170.87	814	2.55	58.8	7.49	348
31	Central African Republic	149.0	52.63	17.75	118.19	888	2.01	47.5	5.21	446



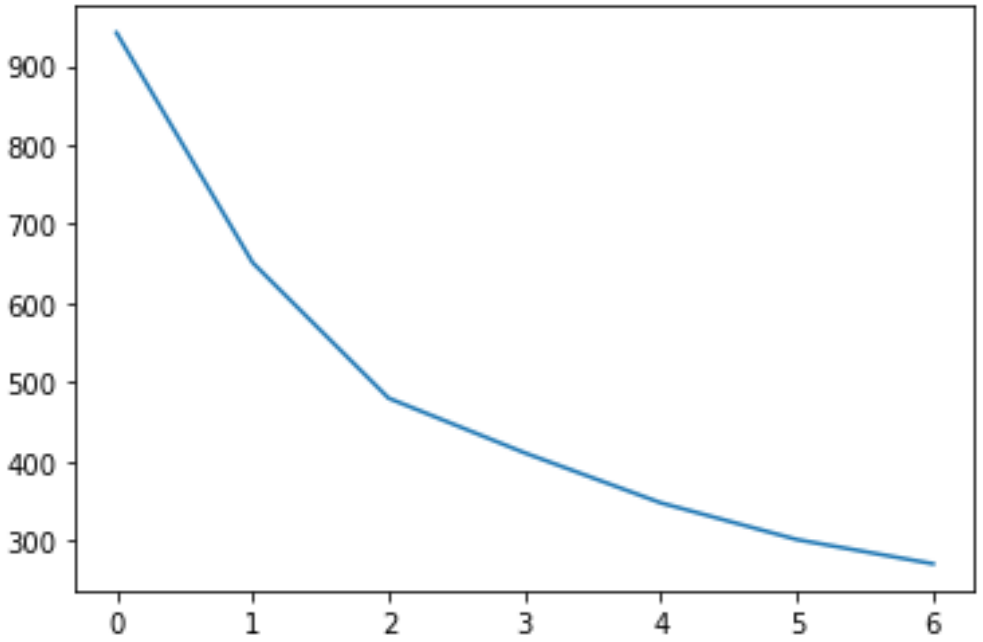
Outliers in Inflation : where inflation > 25 and GDPP is also less

```
In [225]: country_df[country_df.inflation>25].sort_values(by='gdpp',ascending= True).head(10)
```

Out[225]:

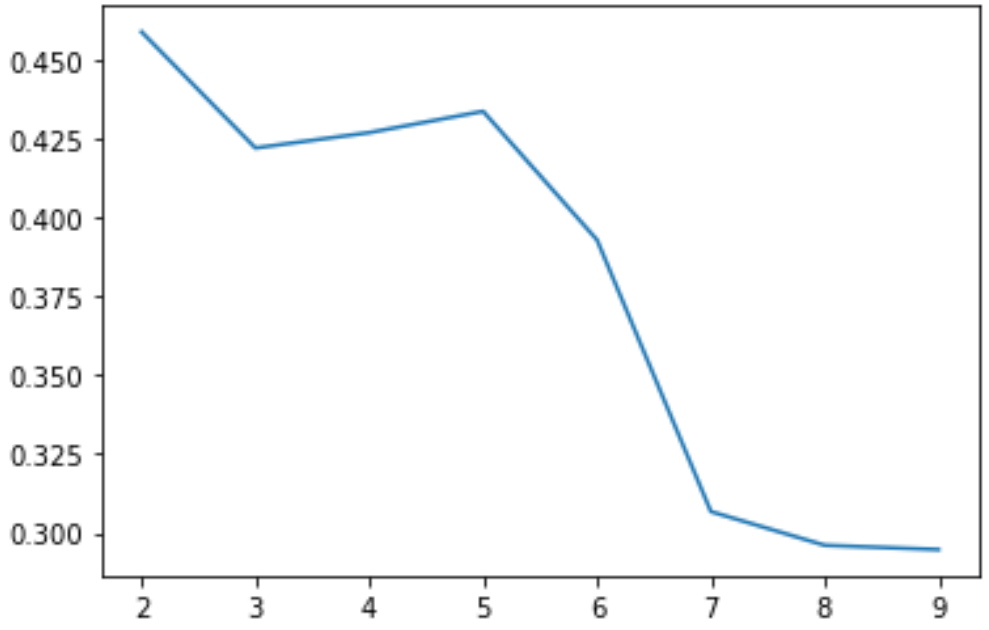
	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
113	Nigeria	130.0	589.49	118.13	405.42	5150	104.0	60.5	5.84	2330
103	Mongolia	26.1	1237.55	144.16	1502.55	7710	39.2	66.2	2.64	2650
149	Timor-Leste	62.6	79.20	328.32	1000.80	1850	26.5	71.1	6.23	3600
163	Venezuela	17.1	3847.50	662.85	2376.00	16500	45.9	75.4	2.47	13500

K- mean Clustering



SSD curve suggests to opt for K= 3

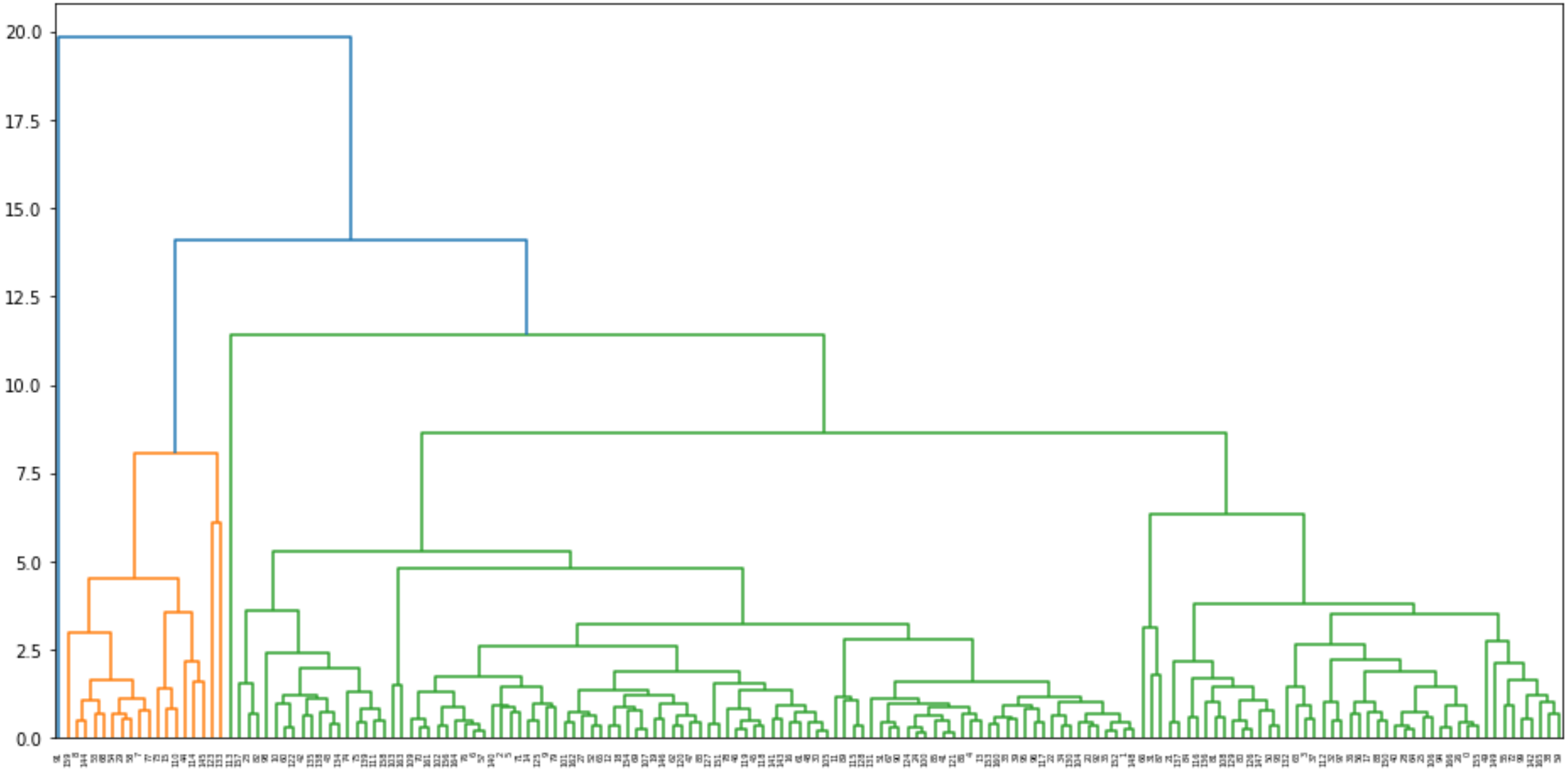
With clusters of 98, 48 & 28 countries each.



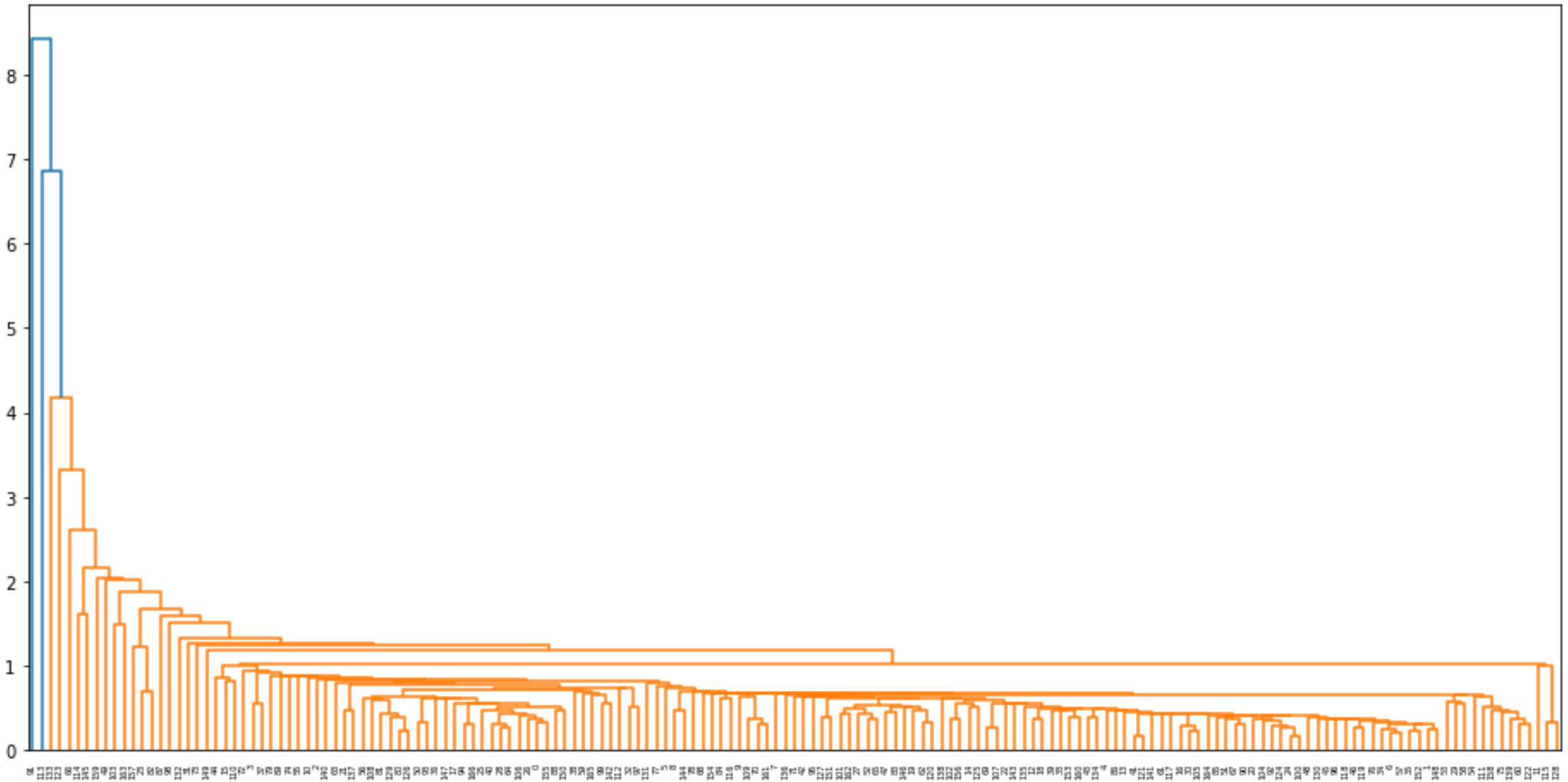
Silhouette curve suggests to opt for k= 2 or 4; But based on the clustering K=3 is the optimal as Business point of view and proceed with that.

Hierarchical Clustering

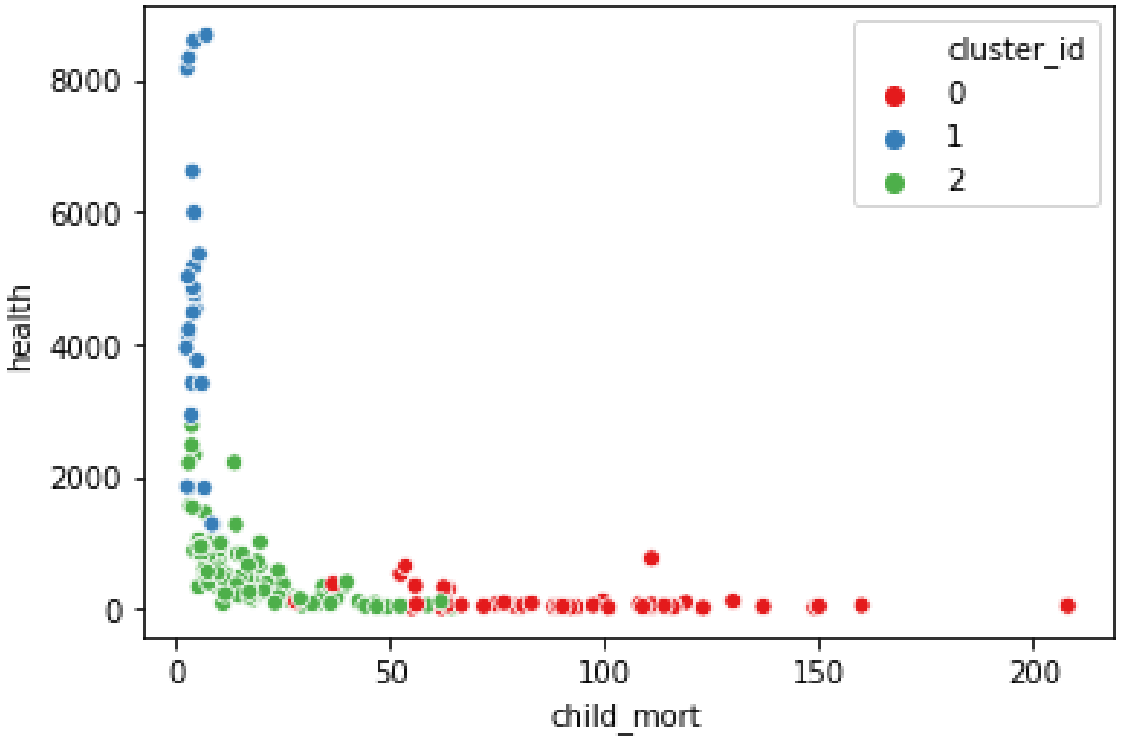
Dendrogram of complete linkage

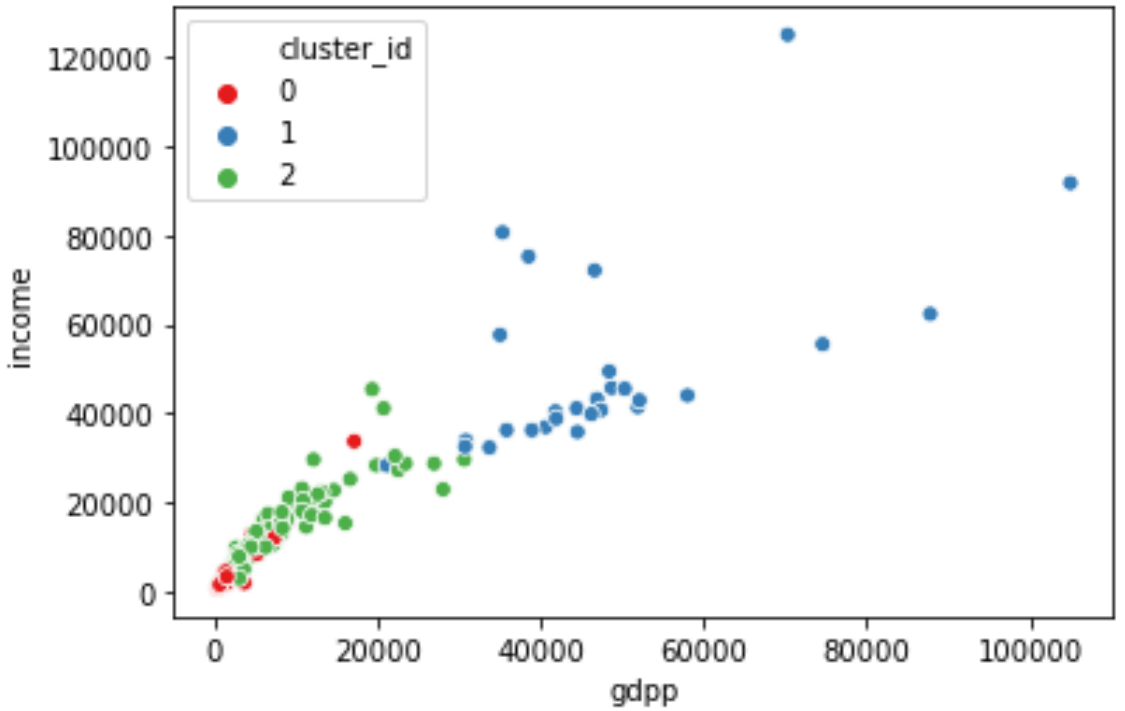


Dendrogram of Single linkage



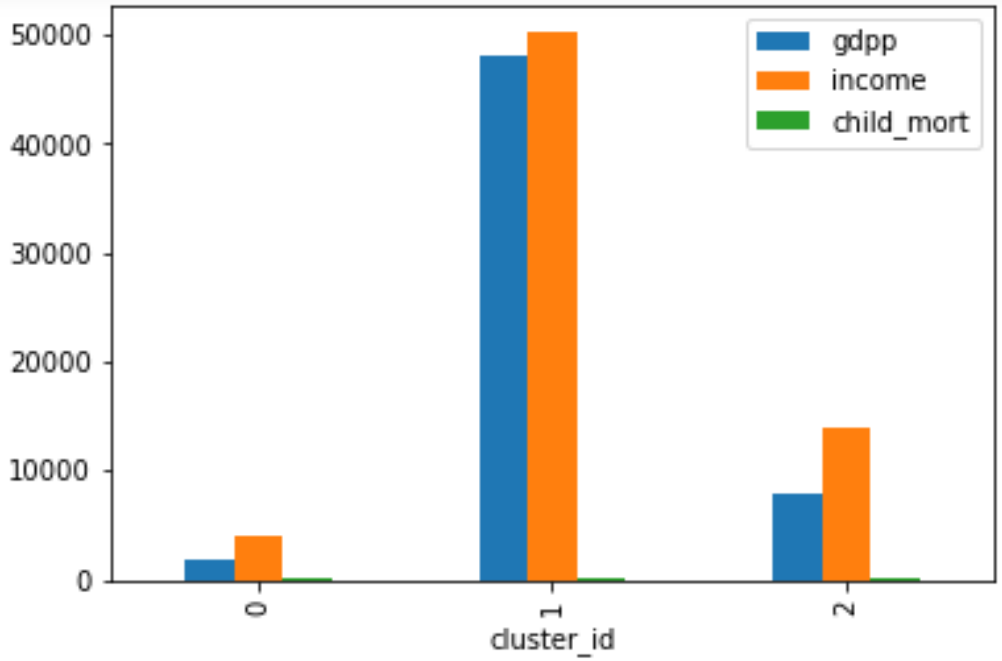
Inferences

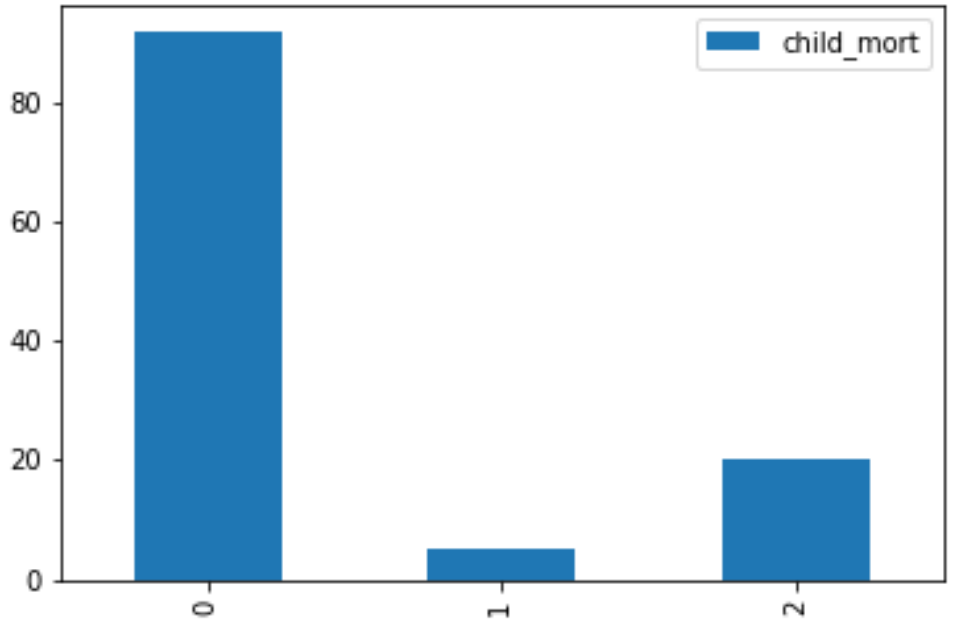




Countries of concern are those where health spending are low and child_mort is high.

Income and GDPP are proportional. And so is health spending wrt GDPP





Inferences

Cluster_id = 0, is the cluster of countries of concern
With low GDPP & Income and high Child_mort.
As can be observed from data below:

cluster_id	gdpp	income	child_mort
0	1909.208333	3897.354167	91.610417
1	48114.285714	50178.571429	5.046429
2	7979.912088	13968.021978	20.357143

Final list of countries that require aid based on Clustering:

```
: c1.sort_values(by = ['gdpp', 'child_mort', 'income'], ascending = [True, False, True]).head(10)
```

:

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	cluster_id
26	Burundi	93.6	20.61	26.80	90.55	764	12.30	57.7	6.26	231	0
88	Liberia	89.3	62.46	38.59	302.80	700	5.47	60.8	5.02	327	0
37	Congo, Dem. Rep.	116.0	137.27	26.42	165.66	609	20.80	57.5	6.54	334	0
112	Niger	123.0	77.26	17.96	170.87	814	2.55	58.8	7.49	348	0
132	Sierra Leone	160.0	67.03	52.27	137.66	1220	17.20	55.0	5.20	399	0
93	Madagascar	62.2	103.25	15.57	177.59	1390	8.79	60.8	4.60	413	0
106	Mozambique	101.0	131.99	21.83	193.58	918	7.64	54.5	5.56	419	0
31	Central African Republic	149.0	52.63	17.75	118.19	888	2.01	47.5	5.21	446	0
94	Malawi	90.5	104.65	30.25	160.19	1030	12.10	53.1	5.31	459	0
50	Eritrea	55.2	23.09	12.82	112.31	1420	11.60	61.7	4.61	482	0


iiit-b
International
Institute of Information
Technology Bangalore
ज्ञानमुत्तमम्


upGrad

Thank you