Clustering & PCA to find
Under-developed Countries in need of aid

Objective: Identify the Under-developed countries in need of aid.

Problem description:

- Based on Socio-economic and health factors of countries, divide the dataset into clusters
- Identify the least developed countries in need of aid

Contents:

- Data preparation, Correlation heat map & it's inferences
- Principal component analysis and Correlation of PCs
- K means clustering with K=4 & K=5, Results- Clusters on PCs as x/y axes , variability across clusters
- Hierarchical clustering with n=4 & n=5, Results- Clusters on PCs as x/y axes
- Final list of countries with K means & Hierarchical clustering
- Conclusion- Countries in direst need of aid

Data- Raw and Prepared

Raw Data

Socio-economic factor data of countries

Country Name	
Child mortality	Life expectancy
Health	Total fertility
Exports	Imports
Income	Inflation

- Dataset size: Rows: 166, Columns : 9

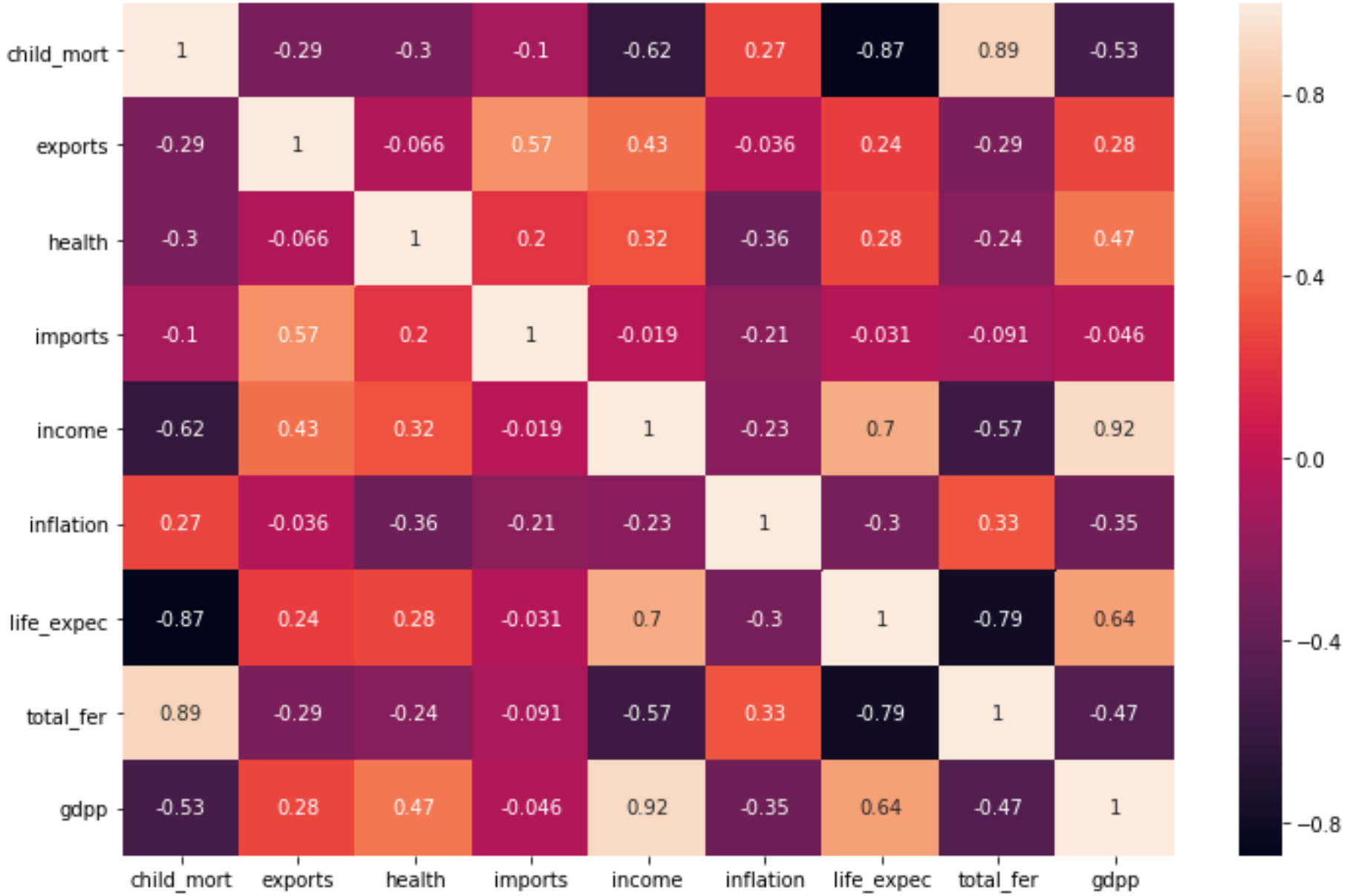
Data Preparation

- No missing values were present hence no treatment was required.
- Outliers were present for all variables, either at higher or lower or both ends.
- Data within +/- 3 standard deviation was retained for all columns.
- Final dataset size: 152 rows
- Scaling was done by StandardScaler() to bring all data on a uniform scale.

Methodology

- Data is prepared by removing outliers and scaling.
- Scaled data is fitted and transformed with PCA to create 4 PCs to reduce dimensionality of data and capture maximum variability.
- The PCA modified data is clustered using K means & Hierarchical clustering.
- Clustering is performed with K=4 & K=5 for both algorithms.
- Variability across clusters is projected on PC1-PC2 and original variable plots.
- Bar graphs explaining variables across clusters are created.
- Final results of both algorithms are summarized, explaining the difference.
- K means results are finally reported.

Correlation coefficient for Numerical variables



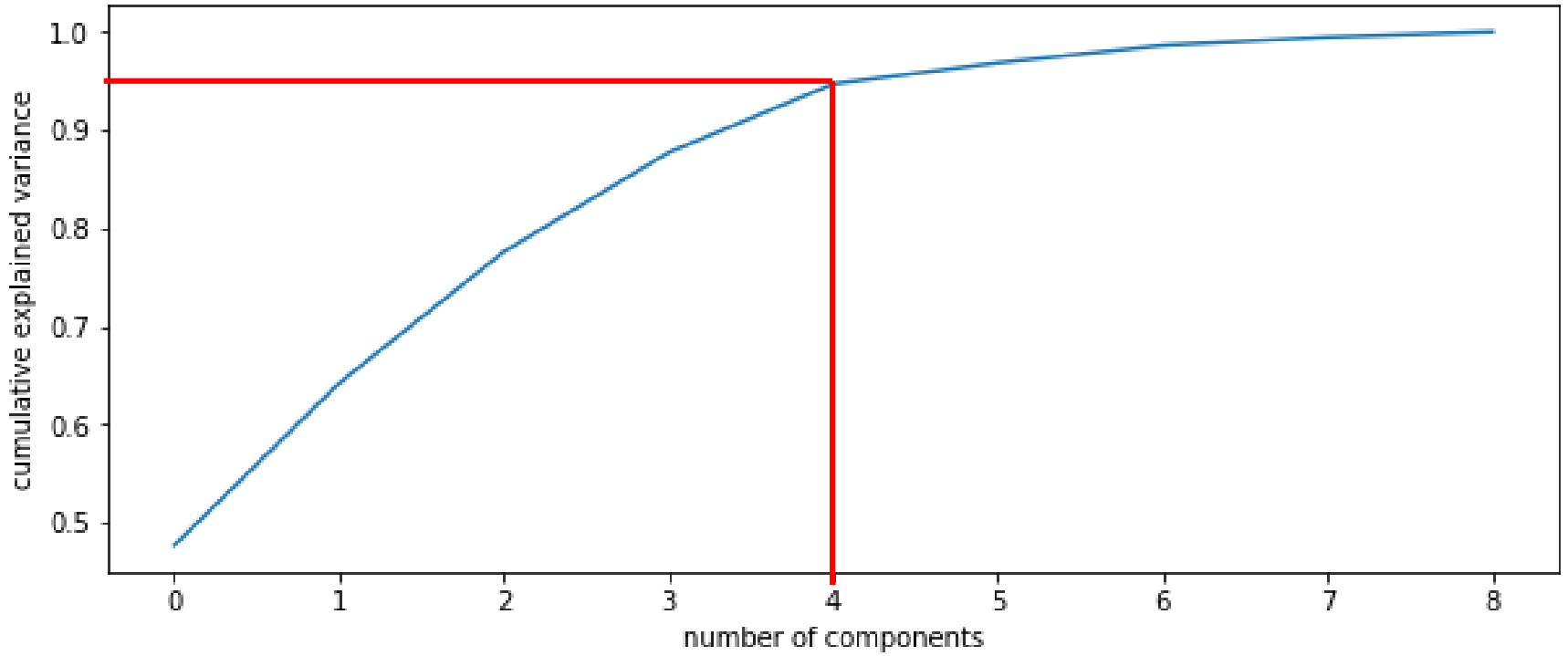
Inferences from correlation heat map

Correlation	Inference
• Very strong positive correlation between child_mort & total fertility • Strong negative correlation b/w child mortality & life expectancy • Moderate negative correlation b/w child mortality & income • Low life expectancy, low income • Strong negative correlation between total fertility & life expectancy • Moderate positive correlation b/w income & exports; high income, high exports • Moderate positive correlation b/w imports & exports; high exports, high imports • GDPP is negatively correlated with total fertility and child mortality • GDPP is positively correlated with income, exports, imports, health.	• Higher fertility is indicative of higher Child mortality • Higher Child mortality drags the life expectancy of a country lower • Lower years of life in a country cause lower per capita income • Higher fertility leads to lower life expectancy, survival rates are lower • Higher income levels indicate development and higher exports from country • Higher exports and higher imports indicate good development levels of country • Low life expectancy leads to lower productivity and lower GDP • High income, exports & imports indicate high GDPP levels.

- Based on this correlation analysis, countries may be classified based on these variables-
- Child mortality, Life expectancy, Health, GDPP, Income

Principal Component Analysis (PCA)

- On the Scaled data, PCA analysis was performed.
- Scree plot-



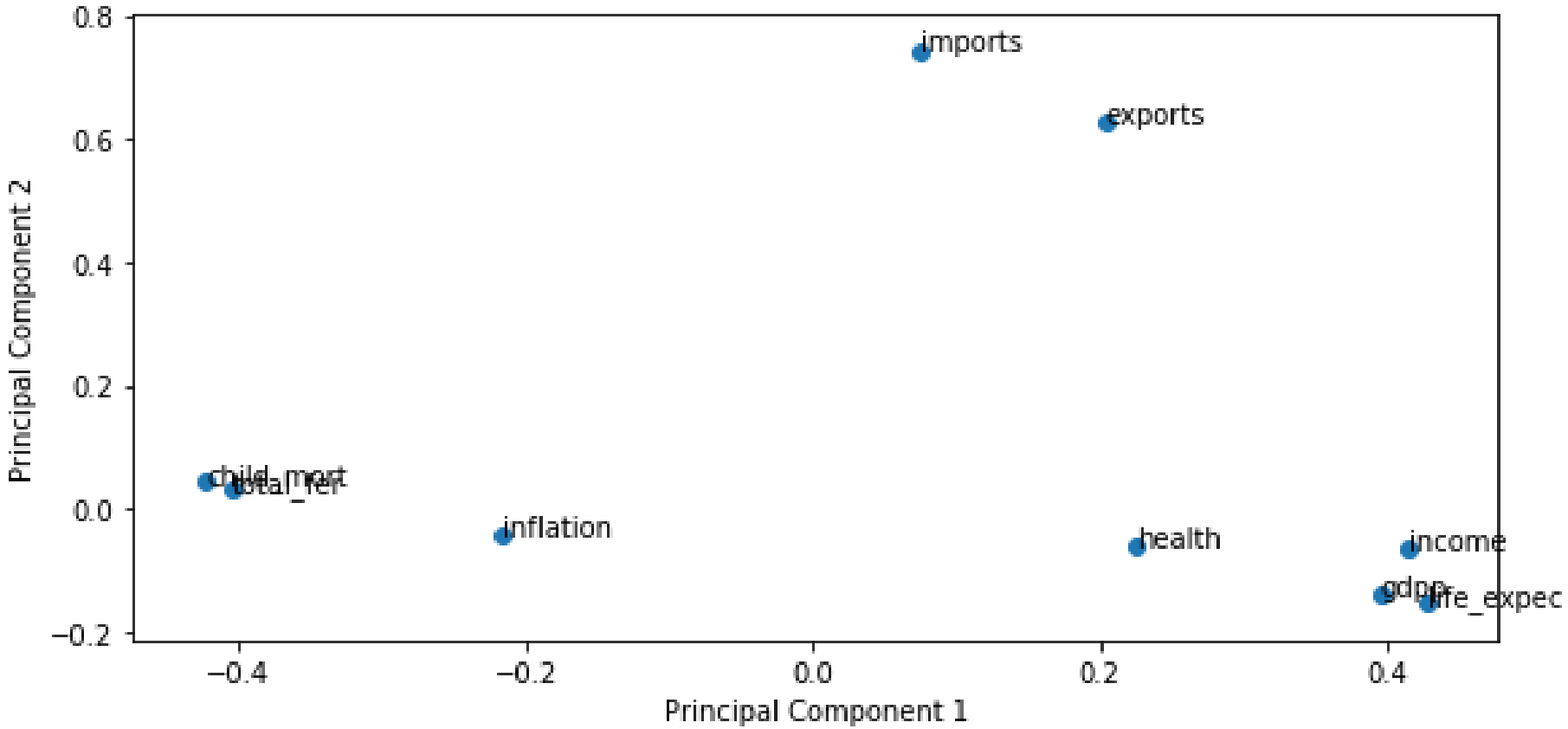
- Four Principal components (PCs) explain 95% variation in the dataset, hence n=4 was selected.

Principal component formation

- For PC selection:
- svd_solver= randomized in PCA, 4 PCs were selected
- The normal PCA method was preferred over Incremental PCA
- As the clustering tendency of PCA modified dataset is better

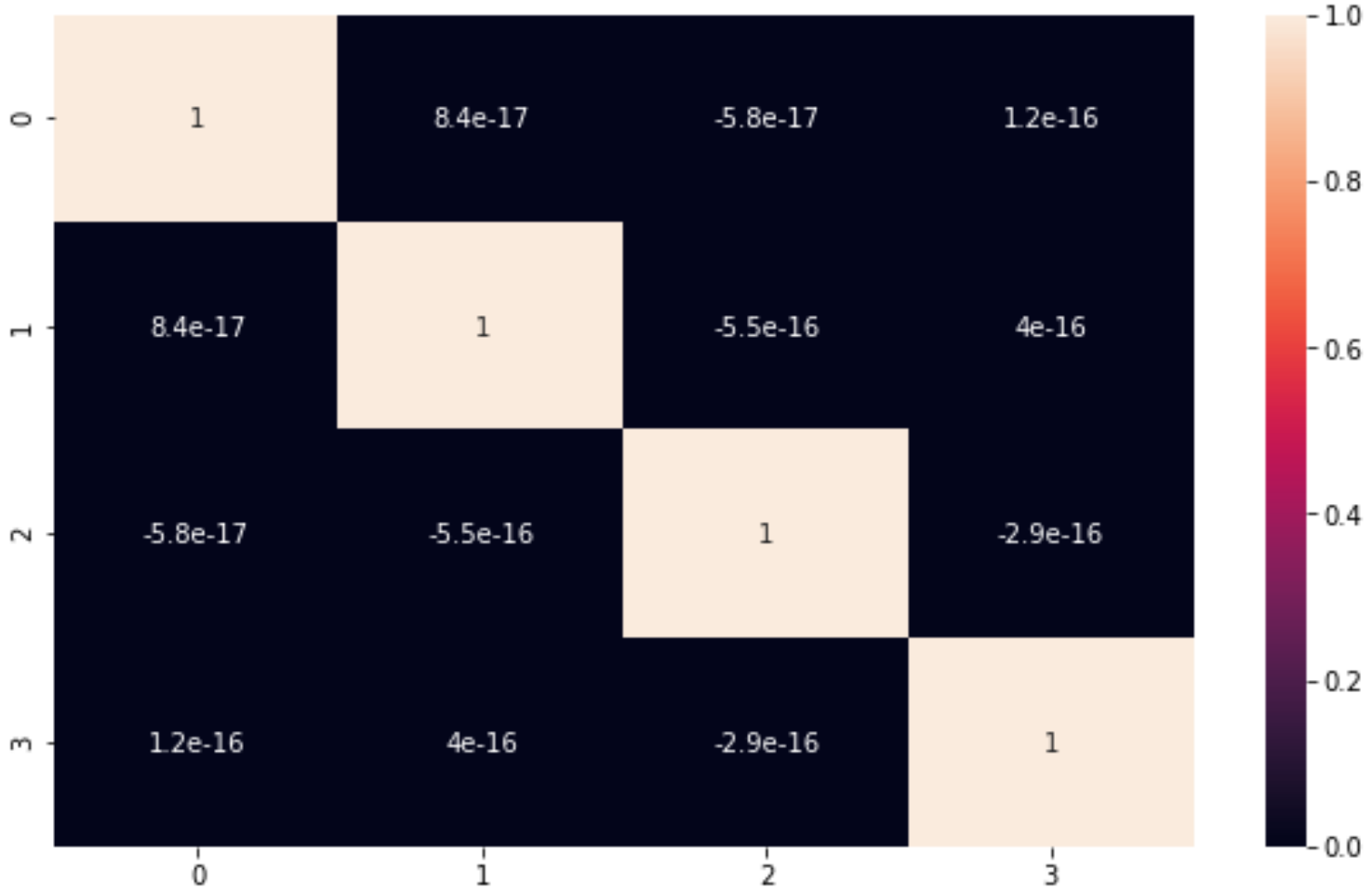
Hopkins stat of PCA modified dataset	
Incremental PCA	Normal PCA with svd
0.68	0.76

PC loading plot



- Loading plot shows the variation explained by principal components PC1 & PC2 for each variable.
- PC1 explains high amount of variability for life expectancy, GDPP, income, child mortality and total fertility.
- PC2 explains higher amount of variability for imports & exports.

Correlation coefficient for PCs



- Minimum and Maximum correlation values are : -2.54*e-16, 2.82*e-16
- This indicates that PCs are independent

PCA shortcomings

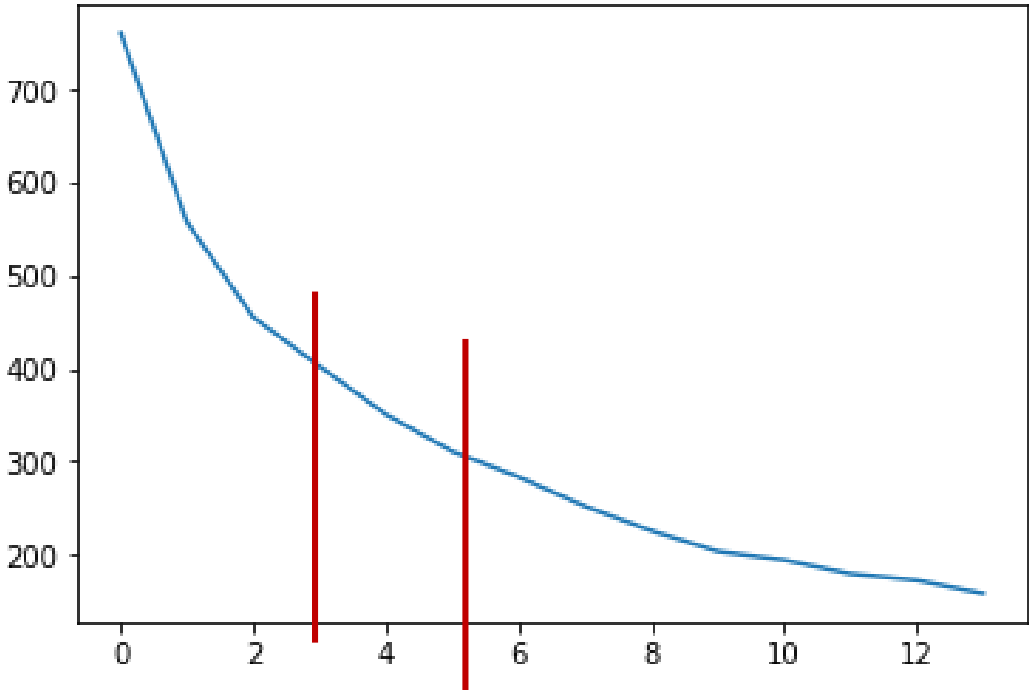
1. PCA concerns with only Linear relationship among variables however there might have been some non-linear relationships as well.
2. PCA makes the perpendicular components to be perpendicular, though in some cases, that might not be the best solution. There might be variations in other directions as well.
3. PCA assumes that columns with low variance are not useful, which might not be true.

Hence we are at risk of losing some information. Still, we have captured 95% of variation.

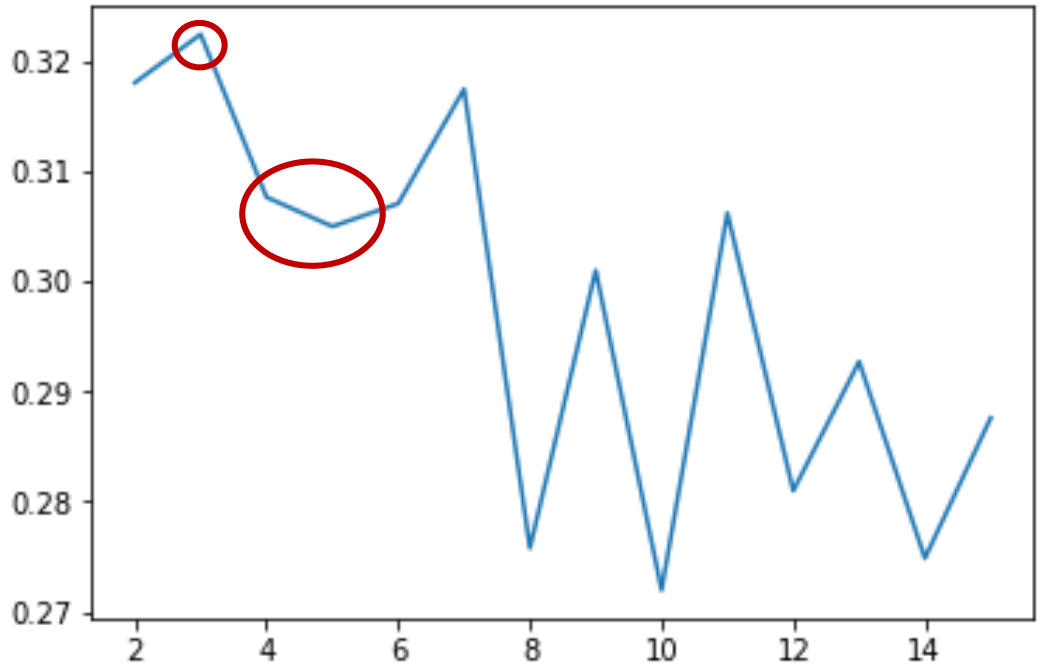
Clustering-k means

- Hopkins statistics is ~0.77, indicating good clustering tendency of PCA modified dataset.

Elbow curve

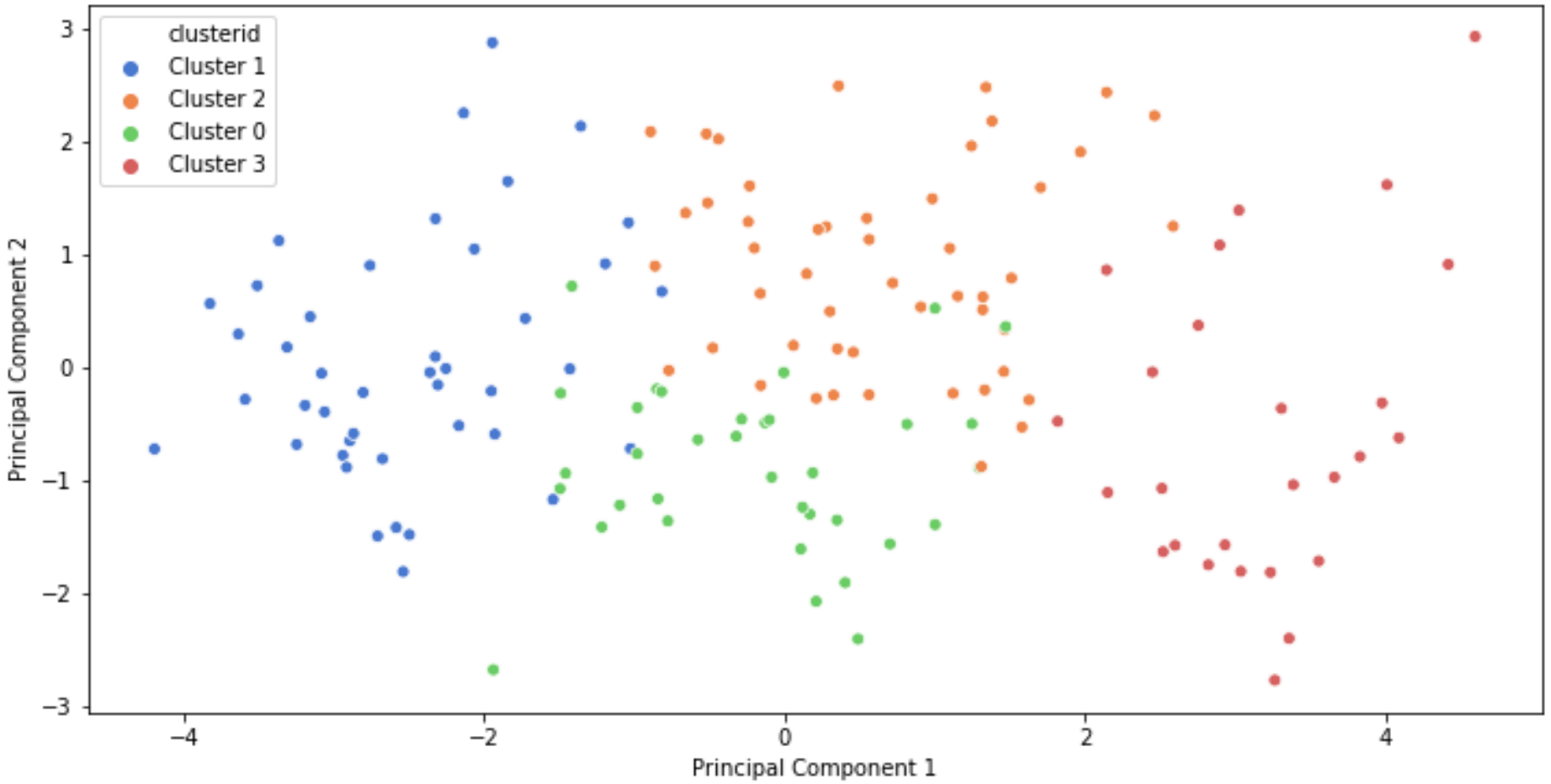


Silhouette scores



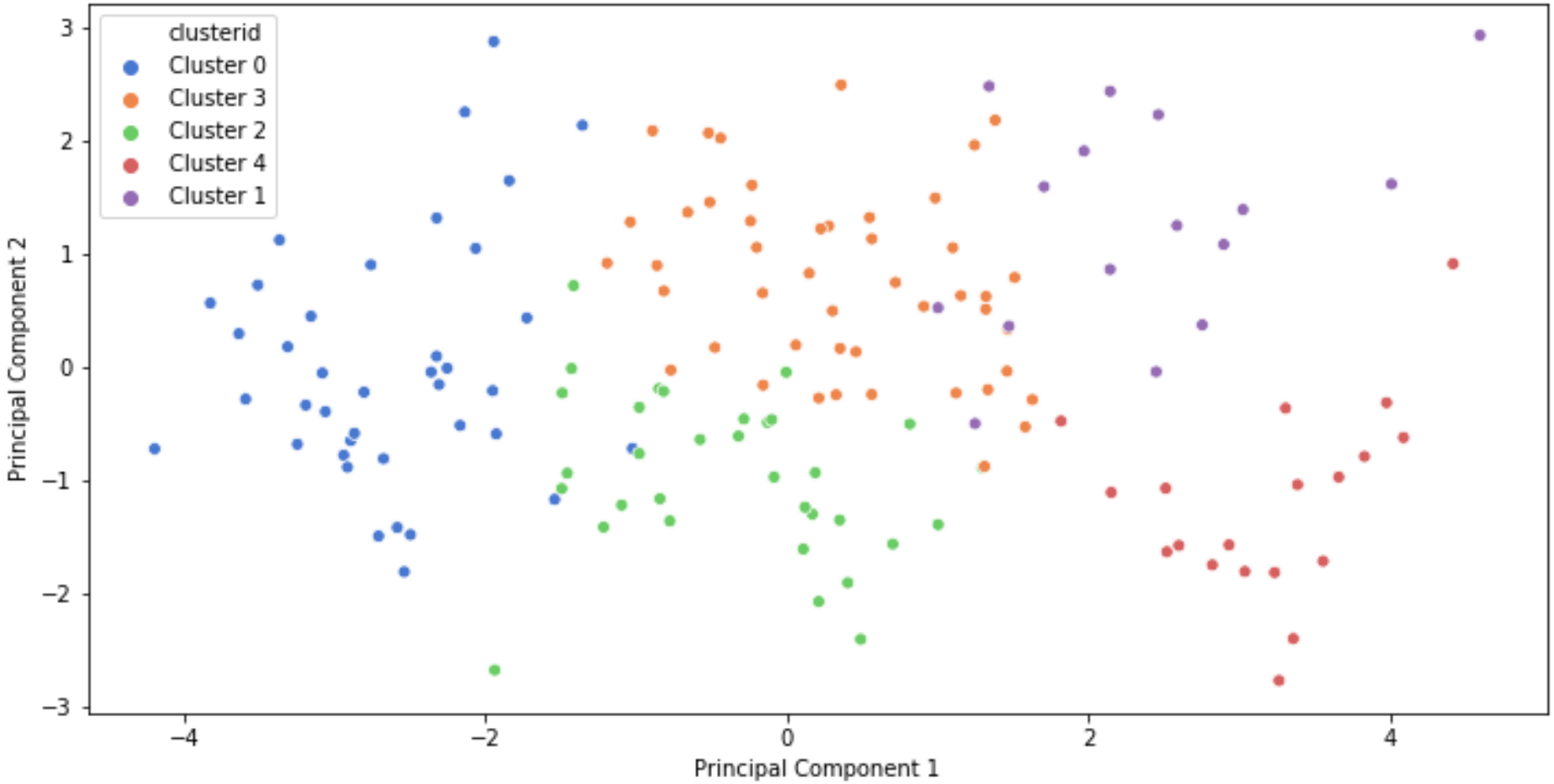
- The elbow curve indicates n=3 to 5 to be optimum no. of clusters.
- At n=3 Silhouette score is maximum at 0.33, however, at n=4 and n=5 ~0.31 i.e. not a large difference.
- Combining the results so elbow curve and Silhouette scores, testing clusters for k=4 and k=5.

K means: Biplot with K=4



- Overlapping clusters are being created with K=4.

K means: Biplot with K=5



- Cluster formation for K=5 is better than K=5.

Number of countries in clusters

K=4

clusterid	
0	35
1	43
2	49
3	26

K=5

clusterid	
0	39
1	16
2	33
3	46
4	19

- K means clustering was performed with K=4 and K=5.
- The number of countries in each cluster for two cases are presented
- K=5 divides the 152 countries (Outlier removed dataset) into smaller clusters.
- Hence it is a reason to prefer K=5 for better understanding of clusters.

K-means: Original variable values across clusters

K=4

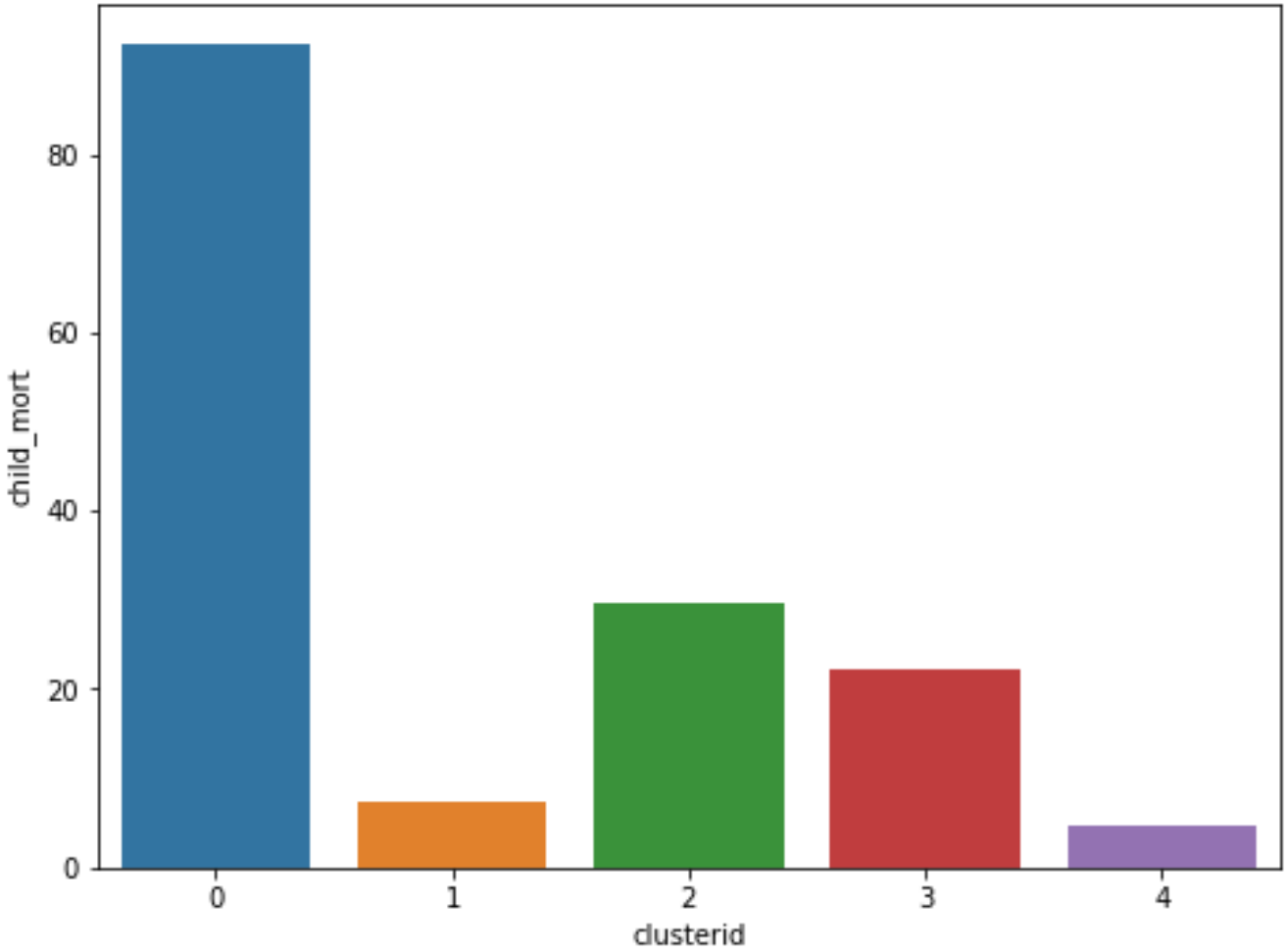
	clusterid	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
0	0	27.60	30.90	5.33	29.11	13657.71	12.02	72.40	2.49	6365.66
1	1	88.91	28.62	6.26	43.68	3521.70	9.71	59.85	5.02	1731.37
2	2	18.15	48.29	6.96	59.89	11741.63	4.00	72.92	2.21	6503.59
3	3	4.83	44.88	9.30	42.57	37442.31	1.81	80.17	1.78	38223.08

K=5

	clusterid	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
0	0	92.29	28.48	6.07	41.74	3226.49	10.03	59.68	5.18	1555.92
1	1	7.23	70.18	6.45	61.42	33718.75	5.16	76.93	1.87	22560.62
2	2	29.52	27.74	5.50	29.13	10903.94	11.70	72.01	2.50	5237.45
3	3	22.30	43.90	7.08	58.75	9959.35	4.29	71.70	2.37	5359.48
4	4	4.68	35.59	10.05	34.98	36578.95	1.27	80.63	1.80	40457.89

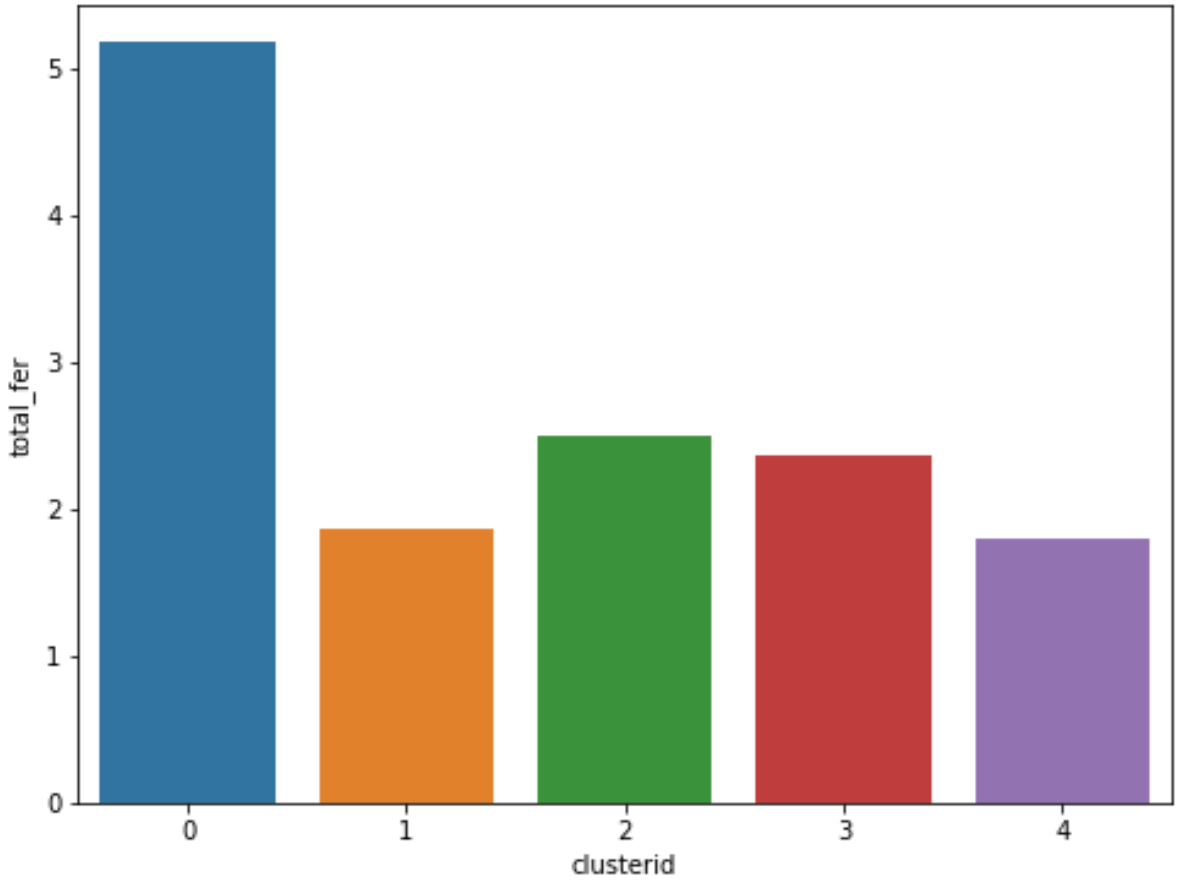
- Variables are distinctly varying across clusters with K=5, with lower and higher ends more extreme than K=4. Hence K=5 was selected for further analysis.

K means: Child mortality across clusters



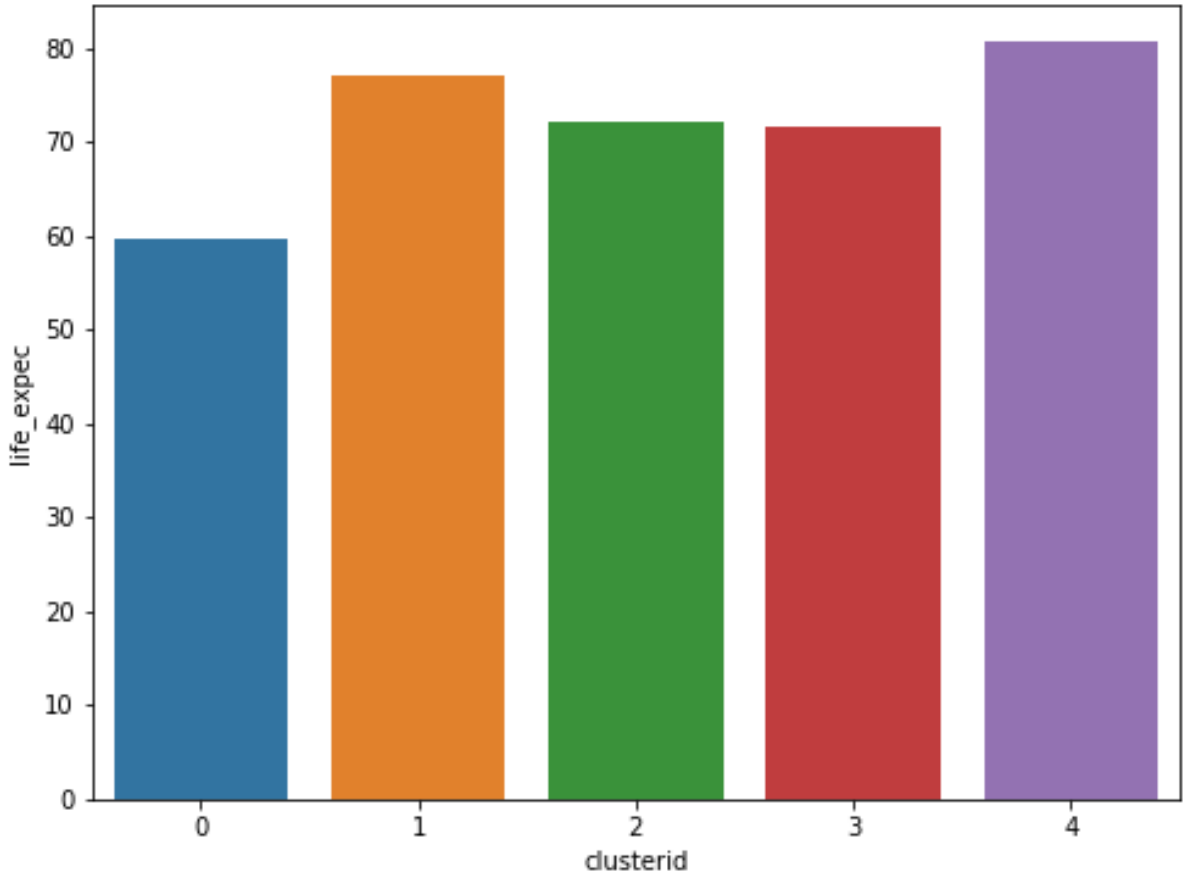
- For Cluster 0, Child mortality is highest, indicating low development for these countries.

K means: Total fertility across clusters



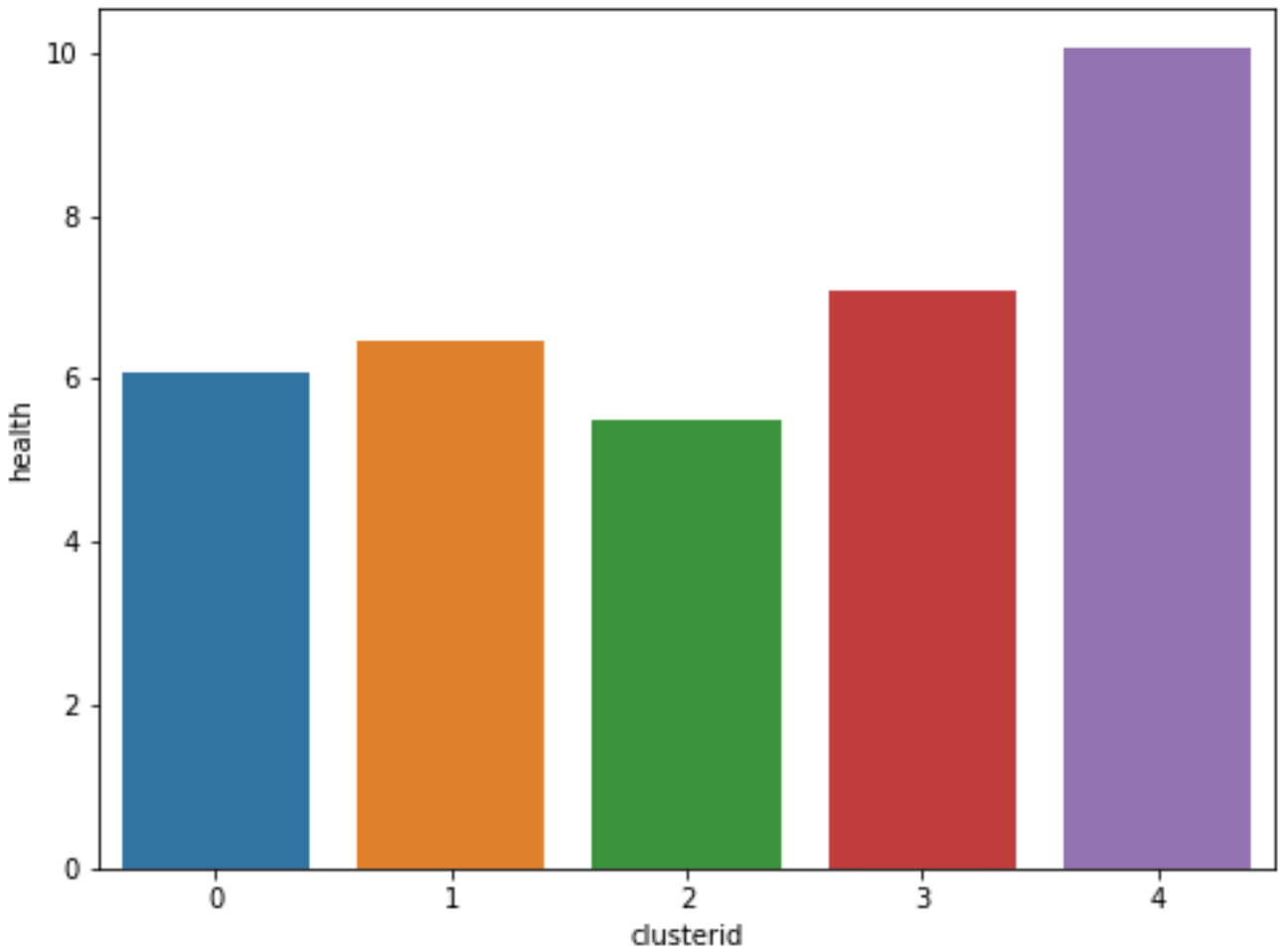
- For Cluster 0, total fertility is highest which has high correlation with child mortality.
- This indicates low socio-economic conditions and low development for these countries.

K means: Life expectancy across clusters



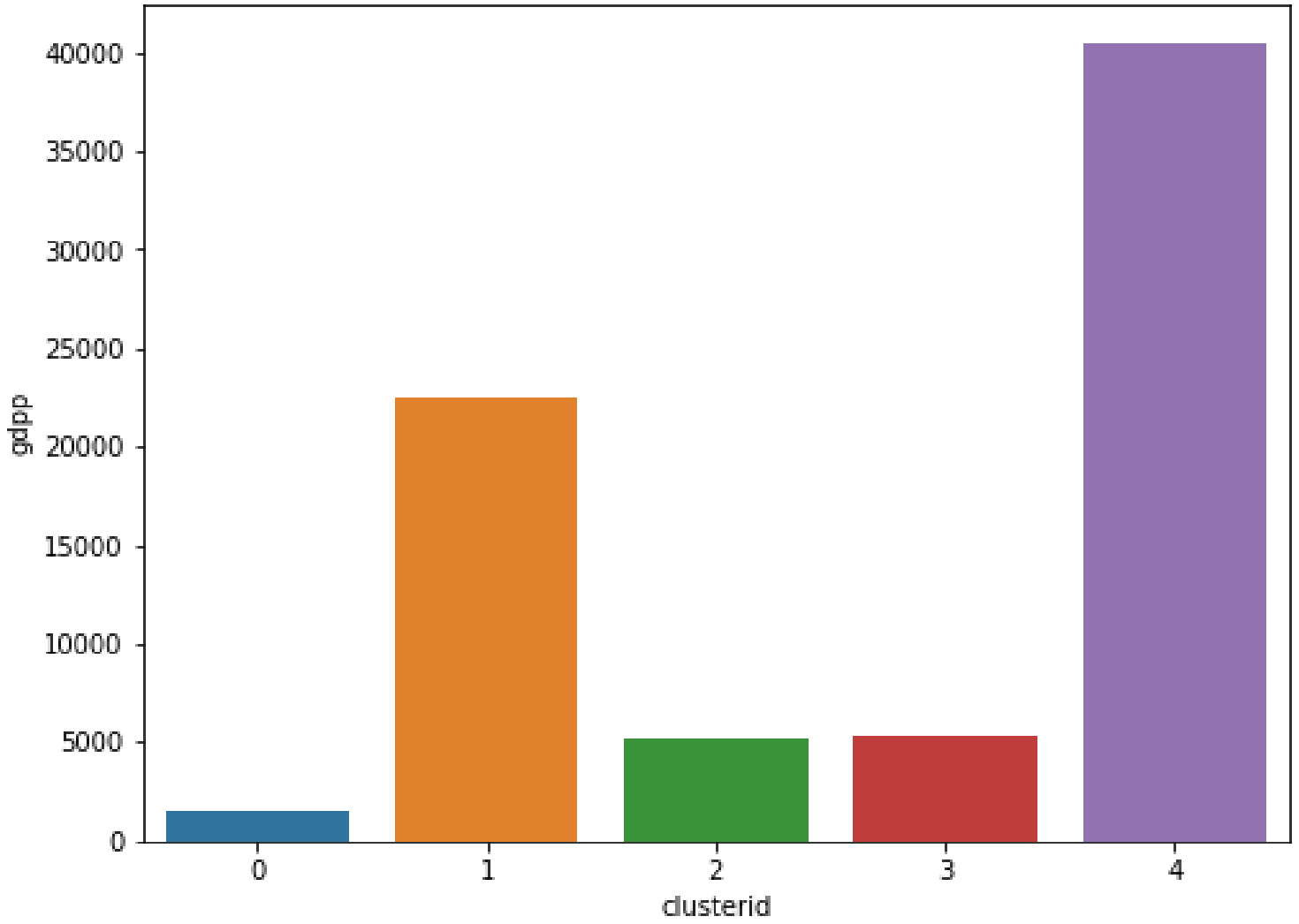
- Life expectancy is lowest for Cluster 0 countries.
- Indicates low socio-economic status.

K means: Health spending across clusters



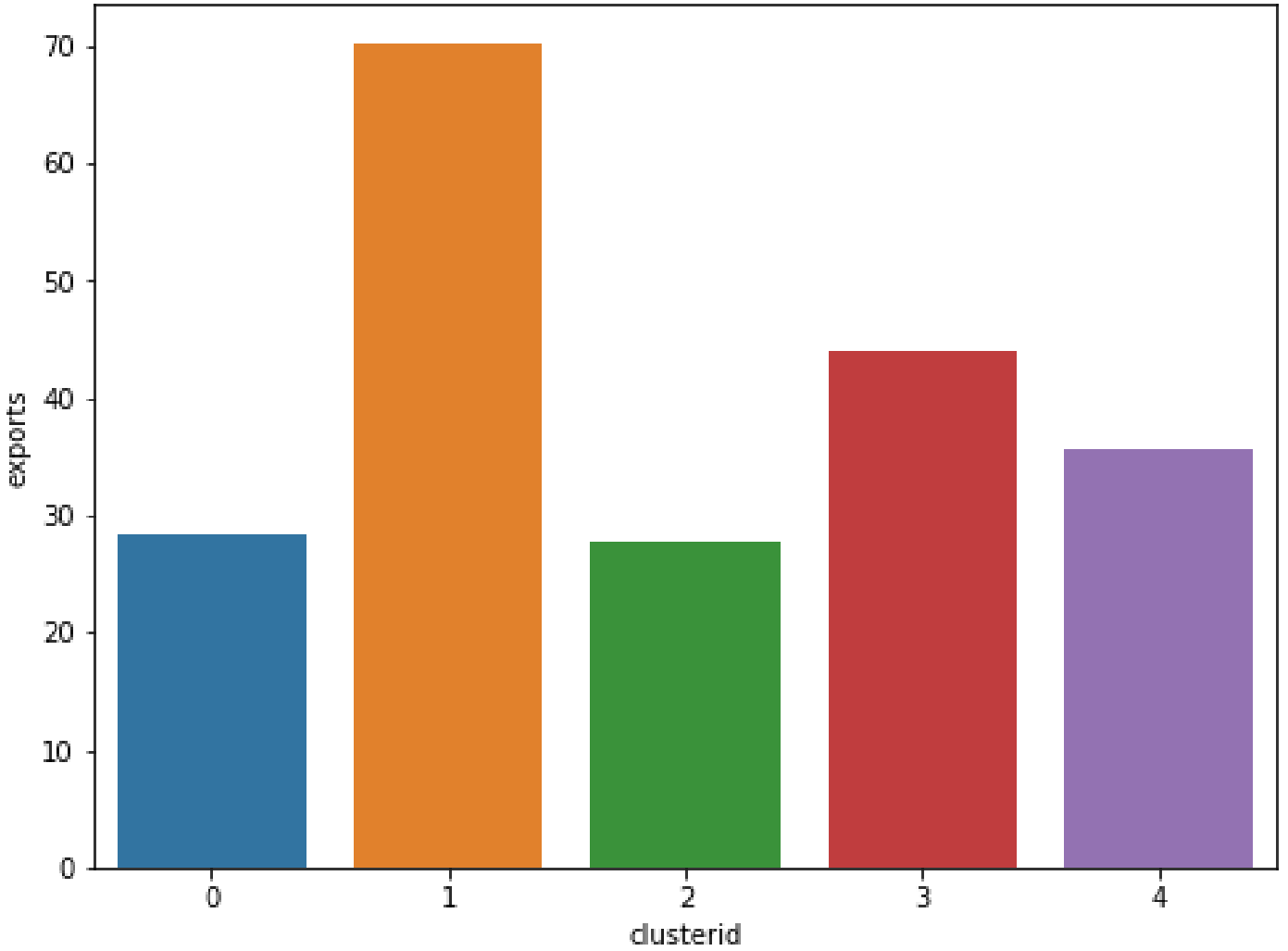
- For Cluster 0 and cluster 2, health spending is low, a sign of undevelopment.

K means: GDPP across clusters



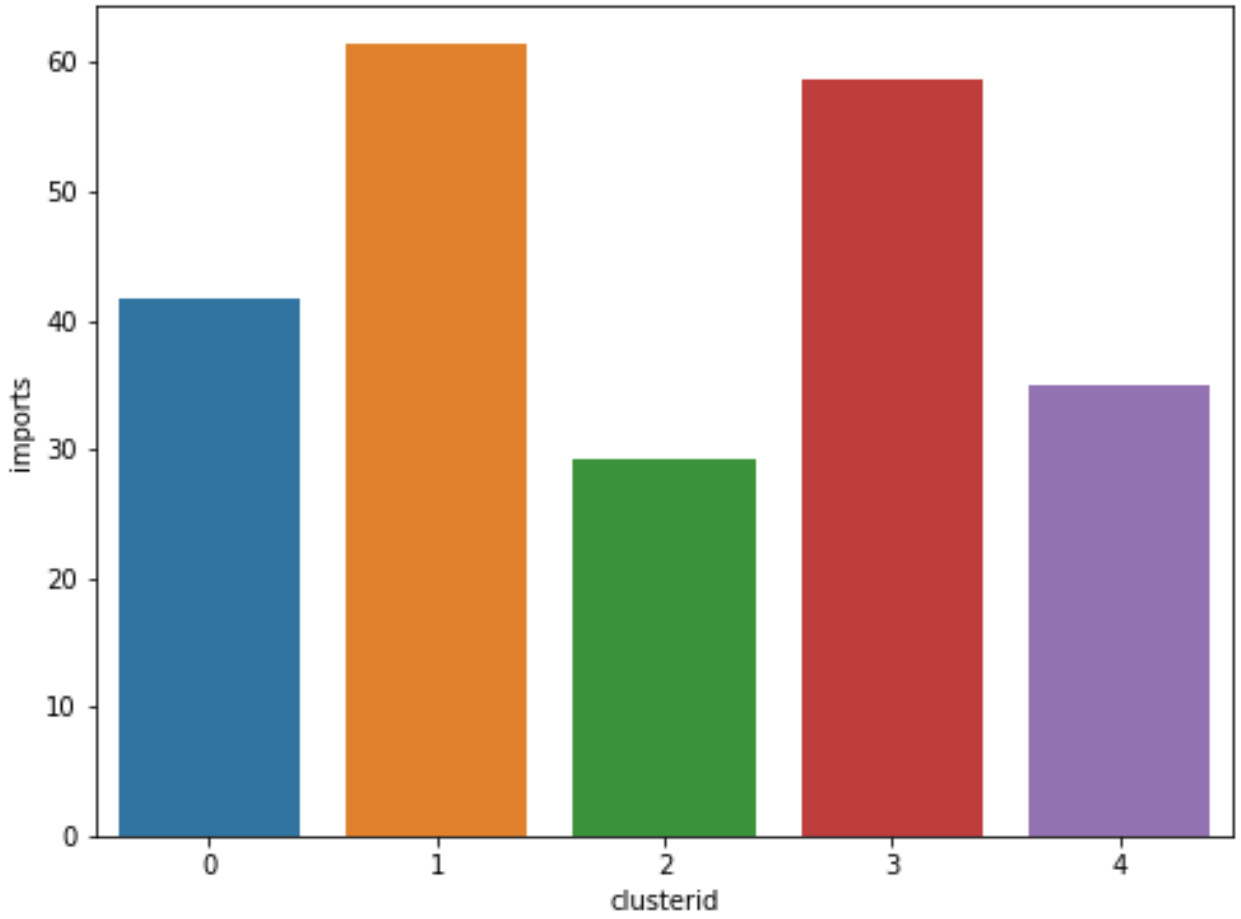
- For Cluster 0, GDPP is lowest, indicating low development for these countries.

K means: Exports across clusters



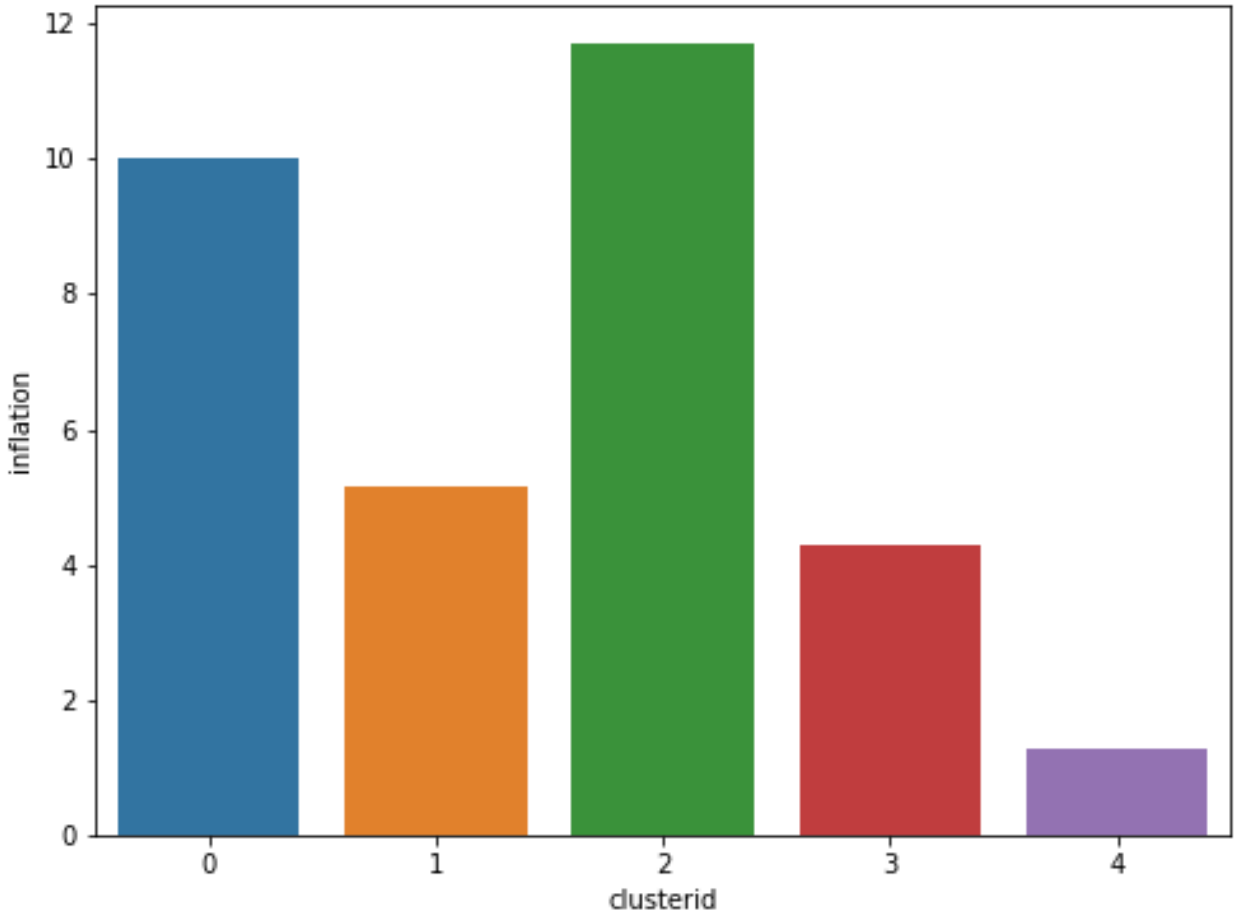
- For Cluster 0 and Cluster 2, Exports are low, indicating low economic activity in these countries.

K means: Imports across clusters



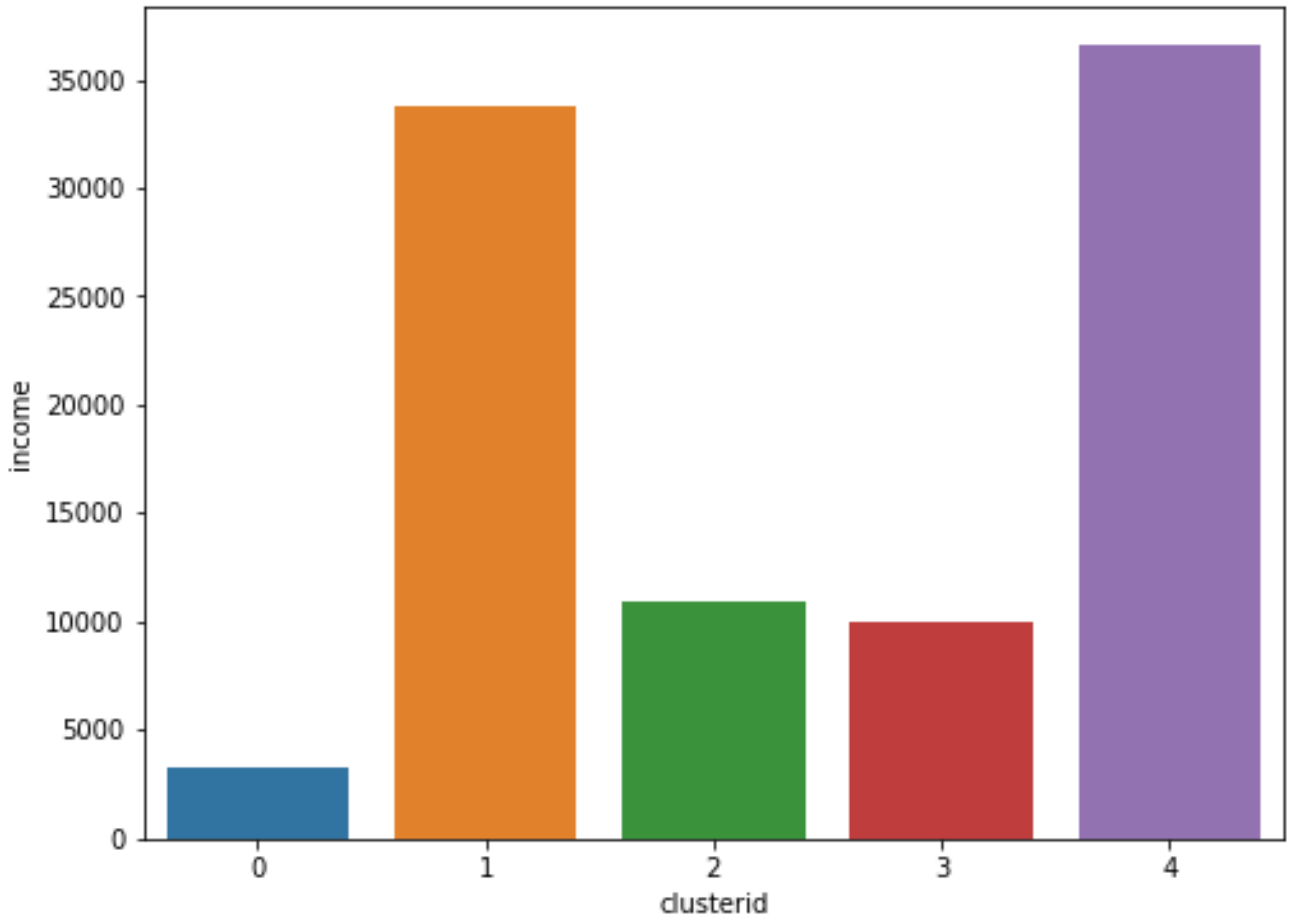
- For Cluster 0 which is collection of low developed countries according to earlier indicators, Imports are significant.
- This indicates country’s dependence on other countries.
- This might be an indicator of low domestic economic activity.

K means: Inflation across clusters



- For Cluster 0 which is collection of low developed countries according to other indicators, Inflation levels are high.
- This might be indication of scarcity of commodities in the country.
- Also, tight economic situation.

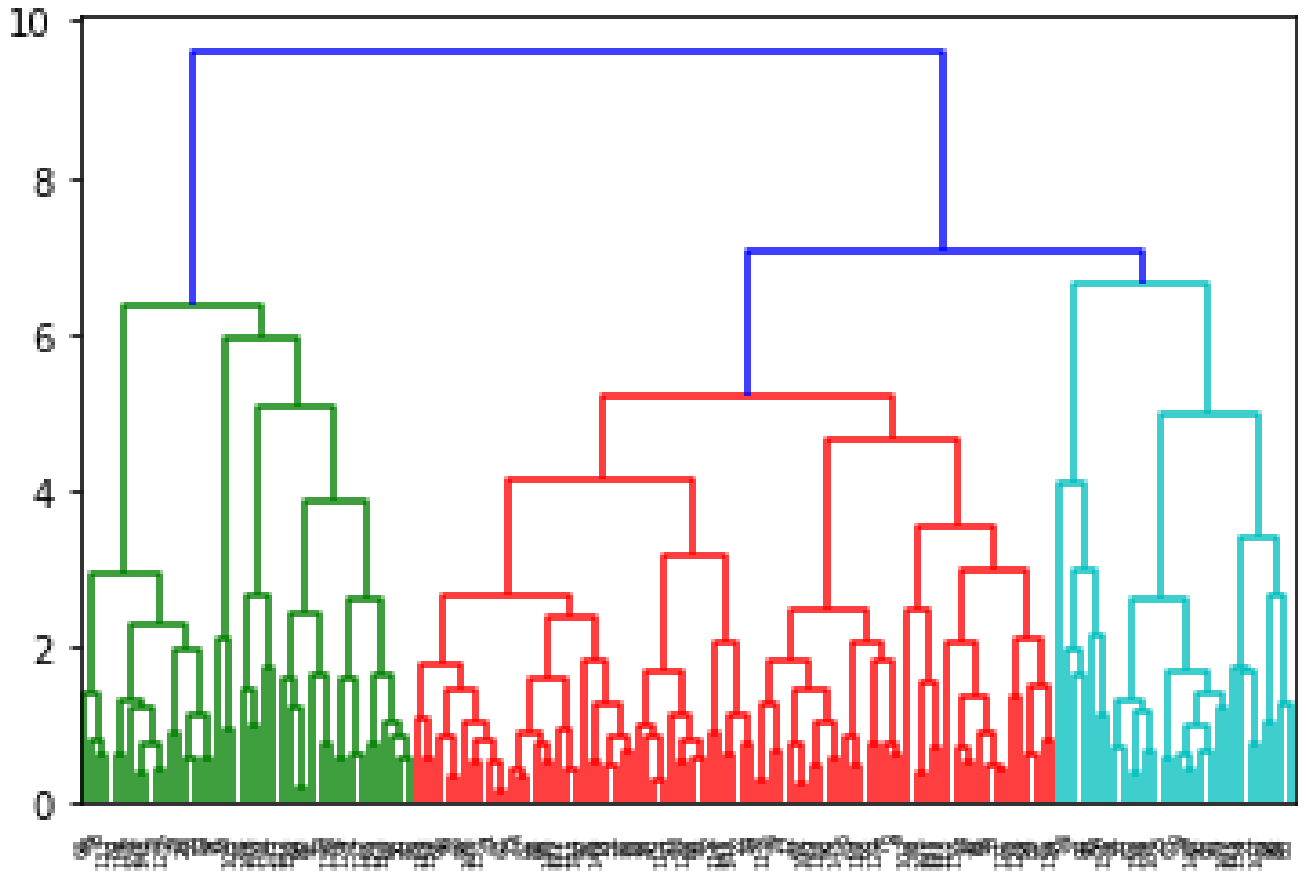
K means: Income across clusters



- For Cluster 0 , Income levels are lowest, indicating low economic development.
- Based on the plots of all indicators, development level of clusters may be written as-
- Bad→ Good: Cluster 0-2-3-1-4

Hierarchical clustering (HC)

Dendrogram: With complete linkage method



Number of countries in clusters

n=4

clusterid	
0	23
1	81
2	7
3	42

n=5

clusterid	
0	23
1	81
2	7
3	17
4	25

- Hierarchical clustering was performed with K=4 and K=5.
- The number of countries in each cluster for two cases are presented
- n=5 divides the 152 countries (Outlier removed dataset) into smaller clusters.
- Hence it is a reason to prefer K=5 for better understanding of clusters.

HC: Original variable values across clusters

n=4

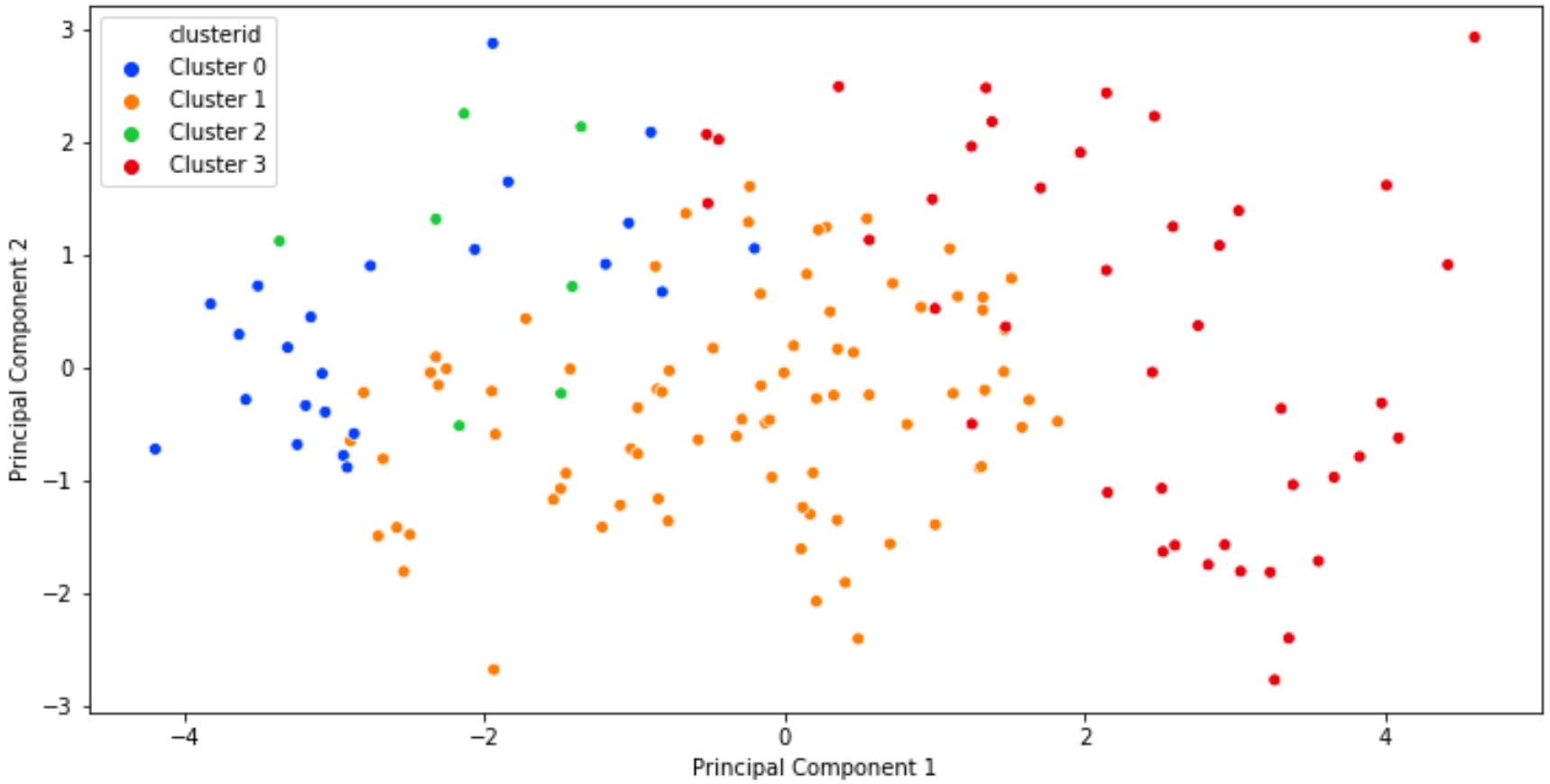
	clusterid	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
0	0	94.91	28.40	7.72	51.51	2384.91	7.34	56.84	5.20	1243.70
1	1	33.17	30.77	6.19	40.35	9351.60	7.21	70.93	2.77	5014.91
2	2	76.77	59.76	4.05	46.81	10814.29	23.76	63.74	4.67	5325.71
3	3	9.44	54.31	7.88	51.38	30720.48	3.85	77.78	1.92	27086.43

n=5

	clusterid	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
0	0	94.91	28.40	7.72	51.51	2384.91	7.34	56.84	5.20	1243.70
1	1	33.17	30.77	6.19	40.35	9351.60	7.21	70.93	2.77	5014.91
2	2	76.77	59.76	4.05	46.81	10814.29	23.76	63.74	4.67	5325.71
3	3	4.15	33.49	10.07	32.78	36858.82	1.39	81.03	1.80	40611.76
4	4	13.04	68.46	6.40	64.03	26546.40	5.53	75.57	2.01	17889.20

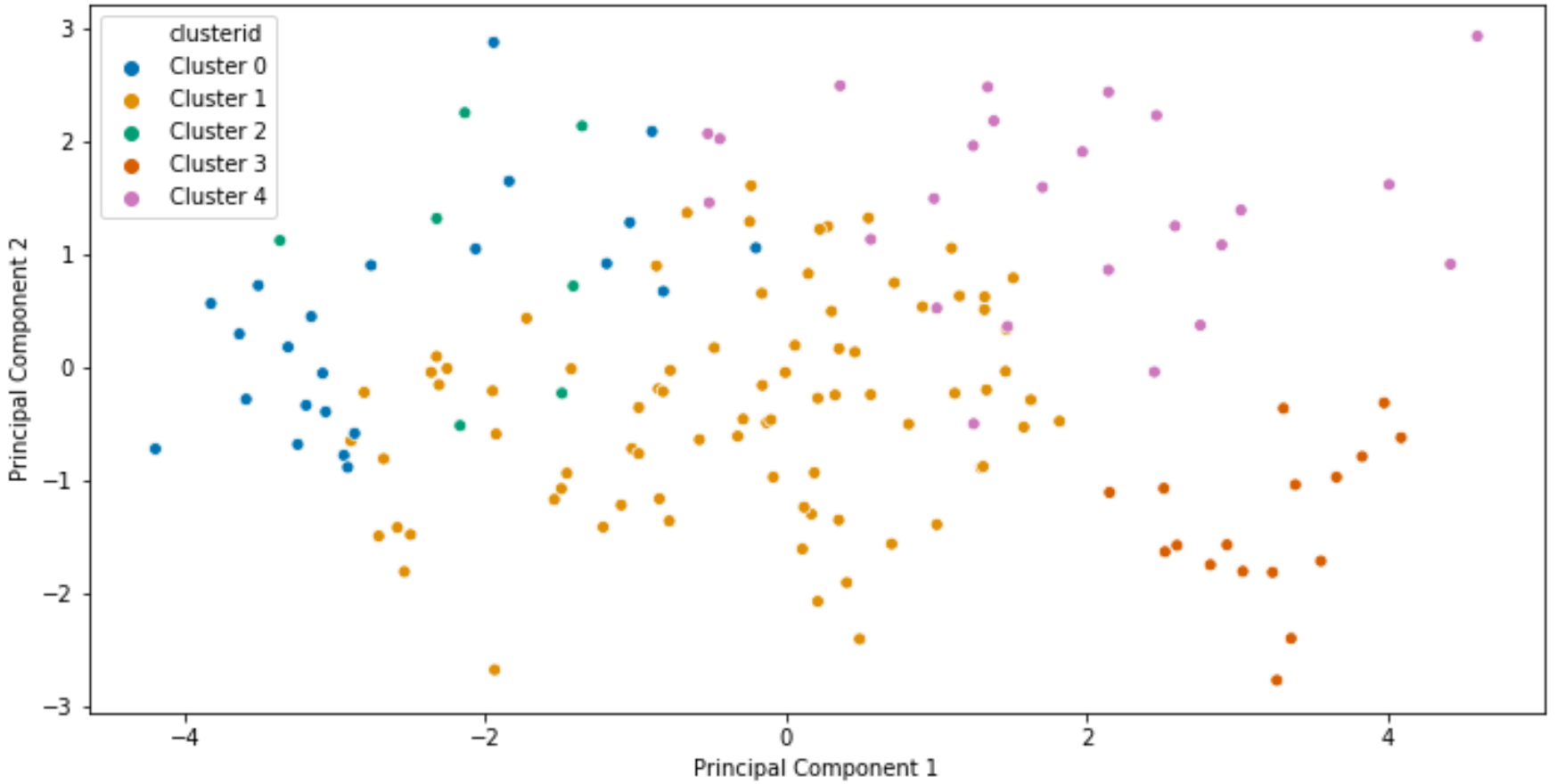
- Variables are distinctly varying across clusters with n=5, with lower and higher ends more extreme than n=4. Above findings are consistent with those of K means.

Hierarchical clustering: Biplot with n=4



Overlapping clusters are forming with n=4.

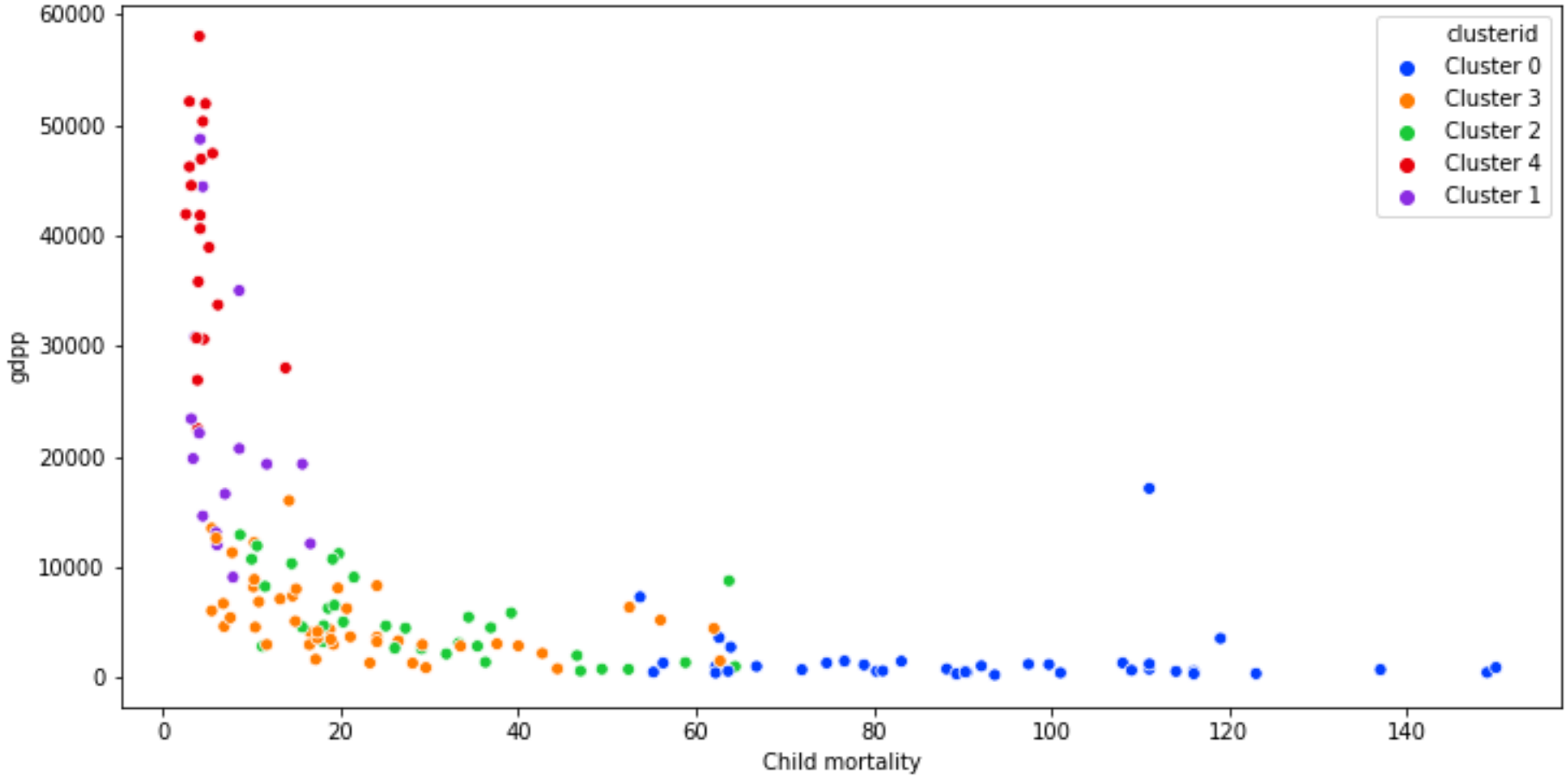
Hierarchical clustering: Biplot with n=5



- Data is divided into more clusters than n=4 but cluster formation is not as distinct as in the case of K means.
- Overlapping clusters are visible. Hence K means algorithm's results are better.

Variable plot- K means clustering

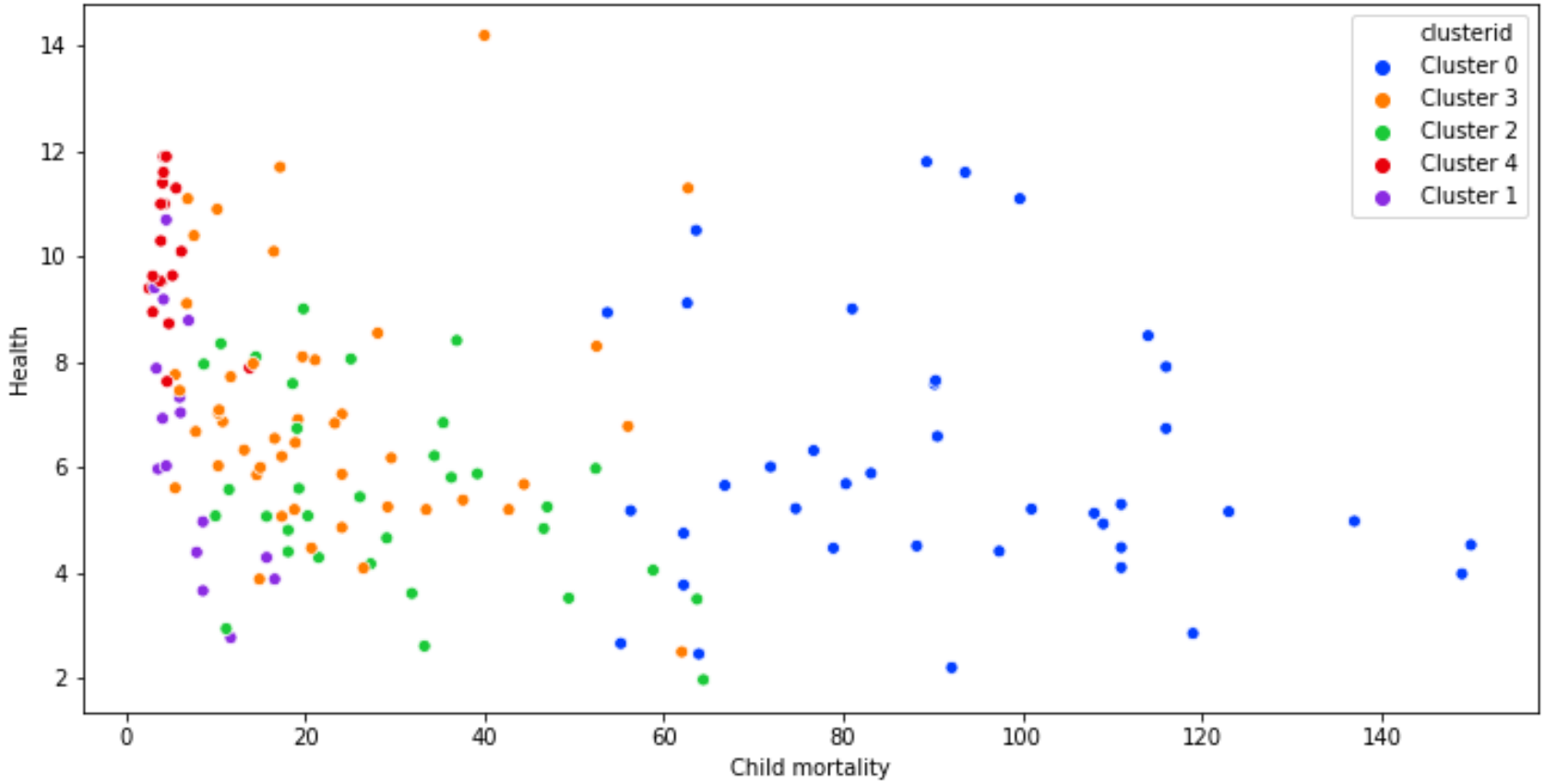
GDPP-Child mortality



- For high Child mortality levels, GDP is low (Cluster 0)
- For low Child mortality levels, GDP is high (Cluster 4)
- The plot explains previously established order of development i.e. Bad to Good
- Cluster 0-2-3-1-4

Variable plot- K means clustering

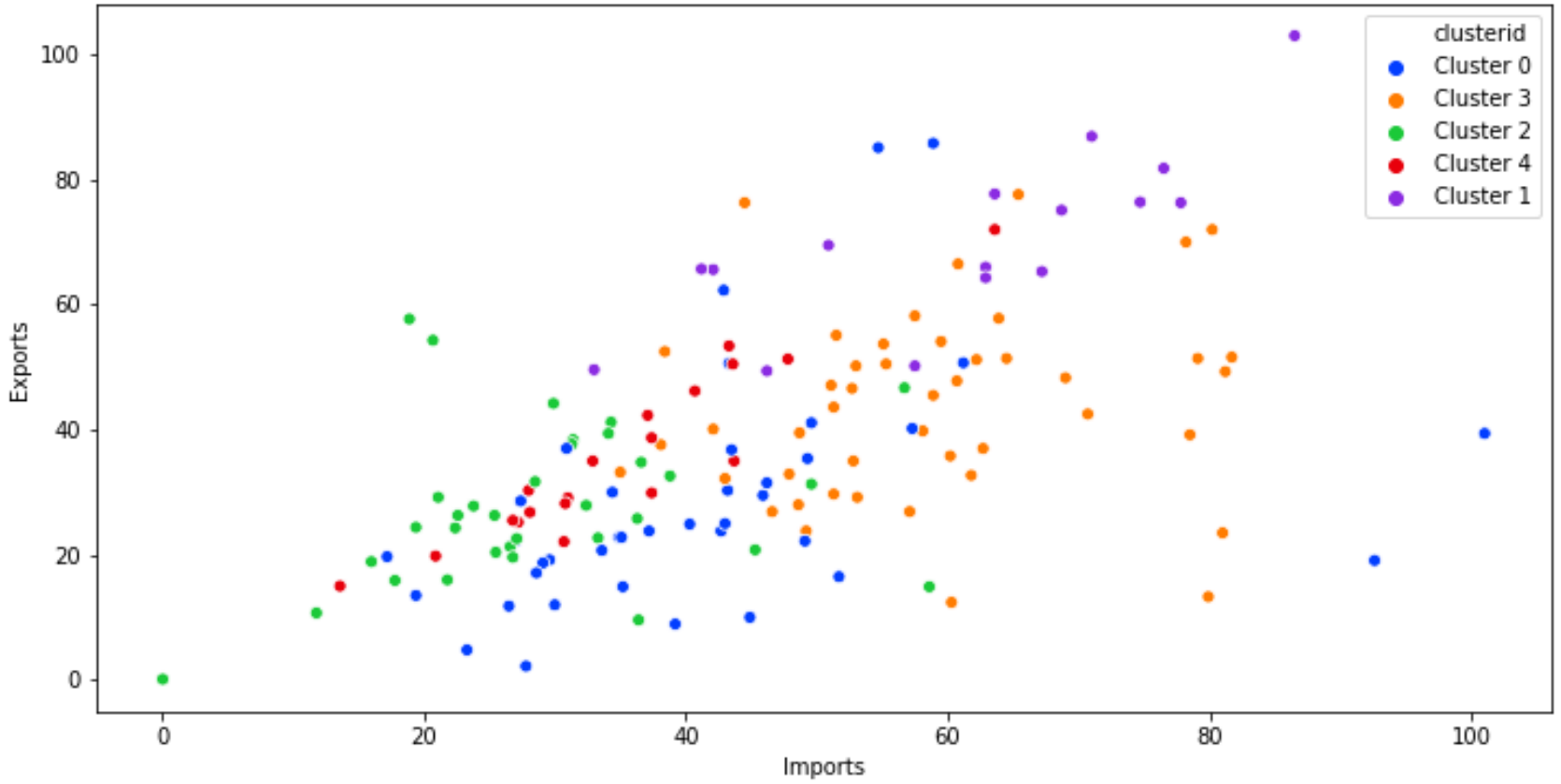
Health spending-Child mortality



- For high Child mortality levels, Health levels are spread across high and low levels for Cluster 0
- But For low Child mortality levels, health levels are clearly high for Cluster 4
- The plot doesn’t clearly distinguish clusters.

Variable plot- K means clustering

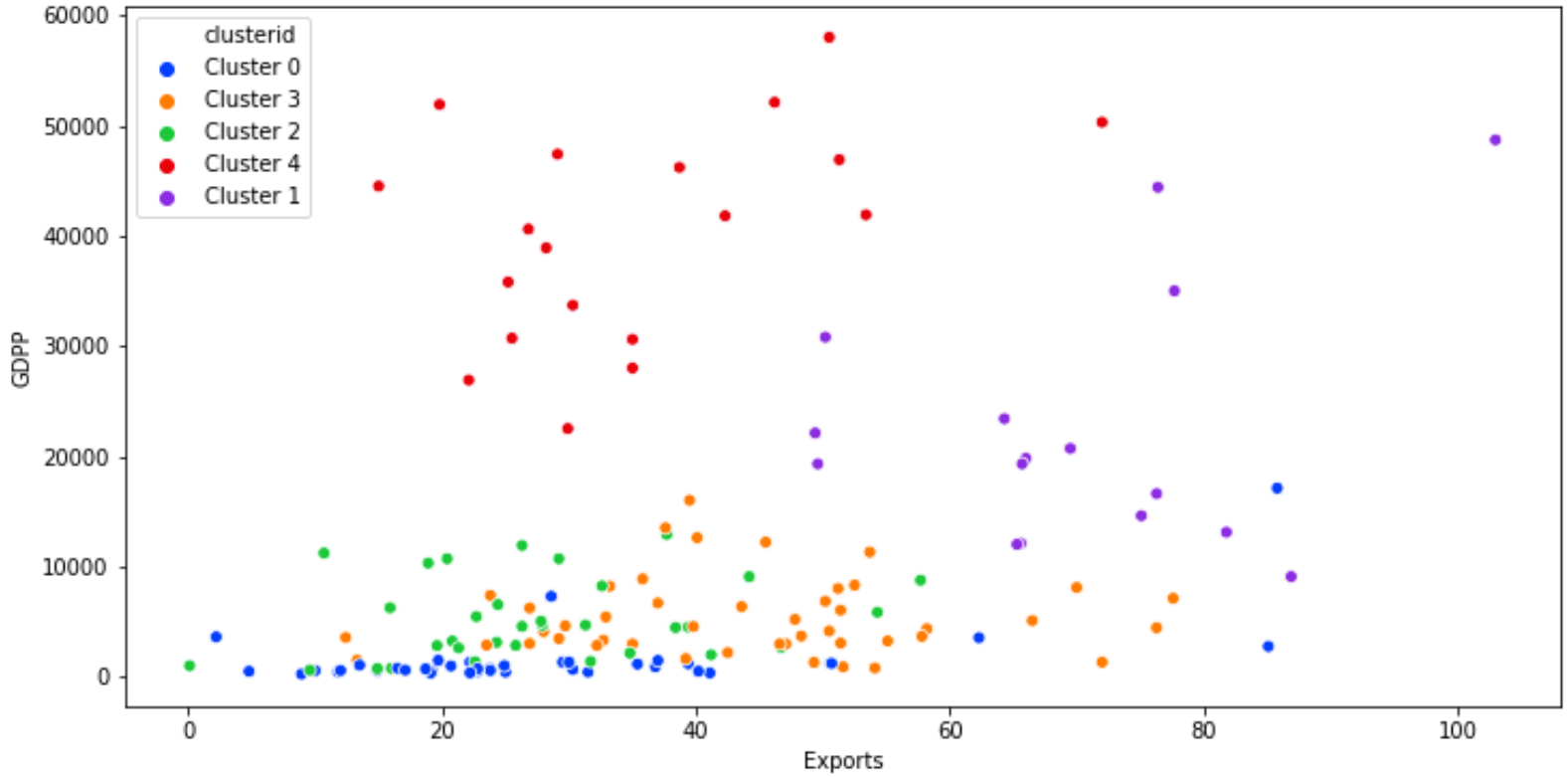
Export-Import



- For cluster 0, exports and imports both are low.
- However this plot doesn't distinguish clusters i.e.
- Low and high developed countries can't be clearly clustered based on exports and import levels.

Variable plot- K means clustering

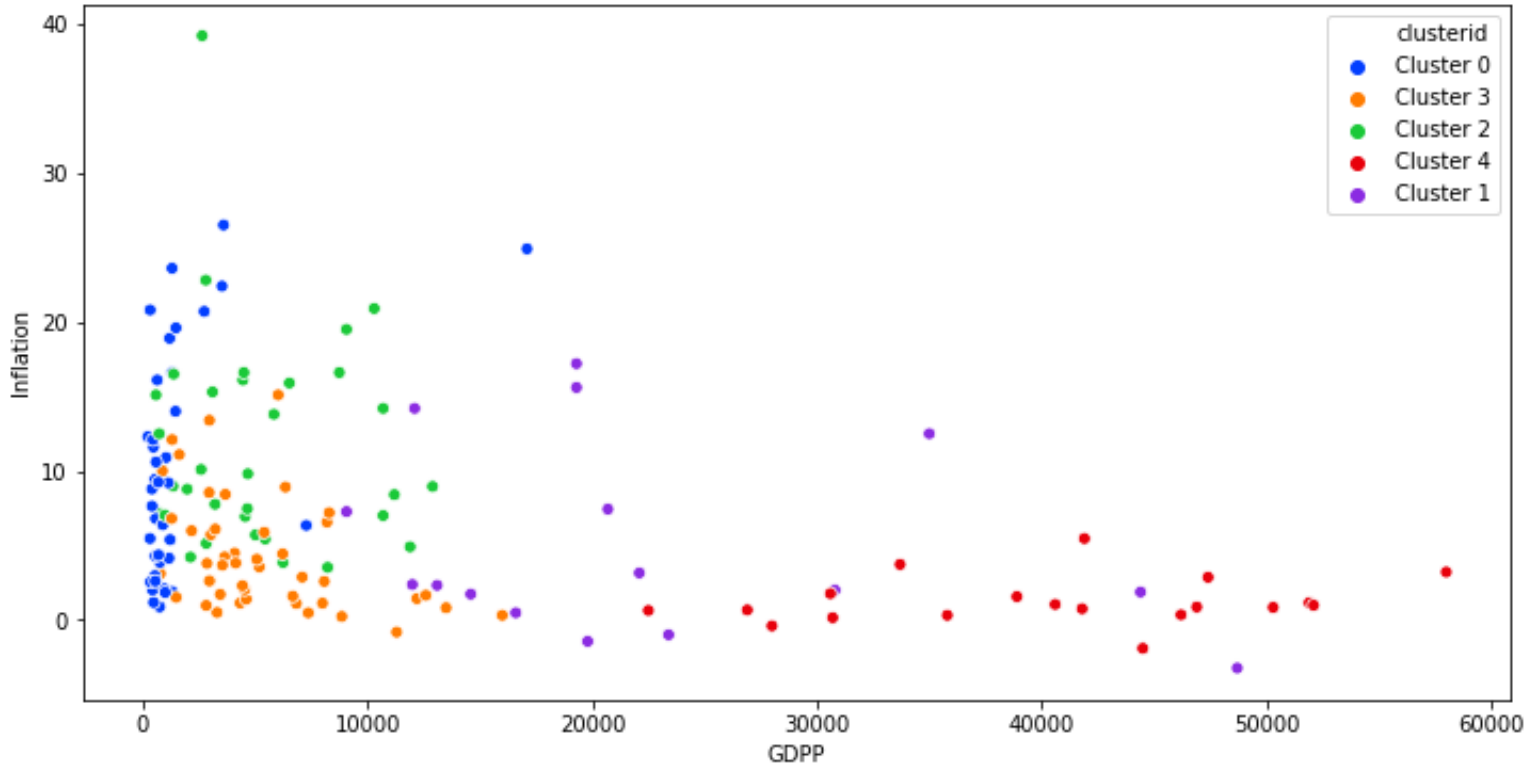
GDPP-Exports



- For cluster 0, exports and GDPP both are low.
- For high developed countries, GDPP is high but exports are low.
- Cluster 2 and cluster 3 also have lower GDP and exports.
- Cluster 1 is developed but lesser than cluster 4
- It has moderate GDP but high exports.

Variable plot- K means clustering

GDPP-Inflation



- For cluster 0, GDPP is low and inflation ranges all values i.e. low and high.
- For high developed countries, GDPP is high and inflation is low due to enough money availability to buy things i.e. scarcity doesn't prevail.
- Cluster 2 and cluster 3 also have lower GDP but moderate inflation, indicating slightly better development levels.

Result: K means clustering

- Based on the mean value of variables across clusters, fining least developed countries from original dataset-
- K means gives list of 8 countries.

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
112	Niger	123.0	22.2	5.16	49.1	814	2.55	58.8	7.49	348
31	Central African Republic	149.0	11.8	3.98	26.5	888	2.01	47.5	5.21	446
106	Mozambique	101.0	31.5	5.21	46.2	918	7.64	54.5	5.56	419
63	Guinea	109.0	30.3	4.93	43.2	1190	16.10	58.0	5.34	648
97	Mali	137.0	22.8	4.98	35.1	1870	4.37	59.5	6.55	708
32	Chad	150.0	36.8	4.53	43.5	1930	6.39	56.5	6.59	897
28	Cameroon	108.0	22.2	5.13	27.0	2660	1.91	57.3	5.11	1310
40	Cote d'Ivoire	111.0	50.6	5.30	43.3	2690	5.39	56.3	5.27	1220

- The least developed countries are sorted in ascending levels of development based on income and life expectancy.
- The development levels are analyzed based on definition of Human development index (HDI) which bases Income, life expectancy and education levels as indicators of development.

Result: Hierarchical clustering

- Based on the mean value of variables across clusters, fining least developed countries from original dataset-
- Hierarchical clustering gives list of 4 countries.

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
31	Central African Republic	149.0	11.8	3.98	26.5	888	2.01	47.5	5.21	446
32	Chad	150.0	36.8	4.53	43.5	1930	6.39	56.5	6.59	897
66	Haiti	208.0	15.3	6.91	64.7	1500	5.45	32.1	3.33	662
106	Mozambique	101.0	31.5	5.21	46.2	918	7.64	54.5	5.56	419

- The least developed countries are sorted in ascending levels of development based on income and life expectancy.
- List of countries is common for K means and Hierarchical clustering.
- However minimum 5 countries has to be reported, hence K means algorithm’s results are final.

Result: Hierarchical clustering

Extracting least developed 10 countries from Undeveloped cluster

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	clusterid
36	Congo, Dem. Rep.	116.0	41.10	7.91	49.6	609	20.80	57.5	6.54	334	0
85	Liberia	89.3	19.10	11.80	92.6	700	5.47	60.8	5.02	327	0
25	Burundi	93.6	8.92	11.60	39.2	764	12.30	57.7	6.26	231	0
107	Niger	123.0	22.20	5.16	49.1	814	2.55	58.8	7.49	348	0
30	Central African Republic	149.0	11.80	3.98	26.5	888	2.01	47.5	5.21	446	0
101	Mozambique	101.0	31.50	5.21	46.2	918	7.64	54.5	5.56	419	0
90	Malawi	90.5	22.80	6.59	34.9	1030	12.10	53.1	5.31	459	0
62	Guinea	109.0	30.30	4.93	43.2	1190	16.10	58.0	5.34	648	0
138	Togo	90.3	40.20	7.65	57.3	1210	1.18	58.7	4.87	488	0
63	Guinea-Bissau	114.0	14.90	8.50	35.2	1390	2.97	55.6	5.05	547	0

- These are almost similar to the ones obtained from K means clustering and binning.

Conclusion: Countries needing developmental aid

- The results are almost similar for K means and Hierarchical clustering.
- The common countries in both the lists are 4-
- However the final selection is subjective to weightage given to different variables.

Niger	Mozambique
Central African Republic	Guinea

- Based on the definition of human development index (HDI), a nation's developmental status is decided by income, life expectancy and education status.
- The case additionally requires us to include health, so K means algorithm’s results obtained after binning are more comprehensive.
- Hence final list of countries in direst need of aid is-

Niger	**Chad**
Central African Republic	**Cameroon**
Mozambique	**Cote d'Ivoire**
Guinea	
Mali	