HELP International

A Case Study to help Nations in need.

Arindam Dey – PGDDS 2019

Synopsis

- **CEO of HELP International needs to prioritize countries that desperately need financial aid.**
- **A recent funding program generated $10 million , that needs to be routed to the countries in need.**
- **Our problem statement is to analyze a set of nations with information related to their socio-economic conditions.**
- **We apply clustering techniques to partition these countries that are similar socio-economically.**

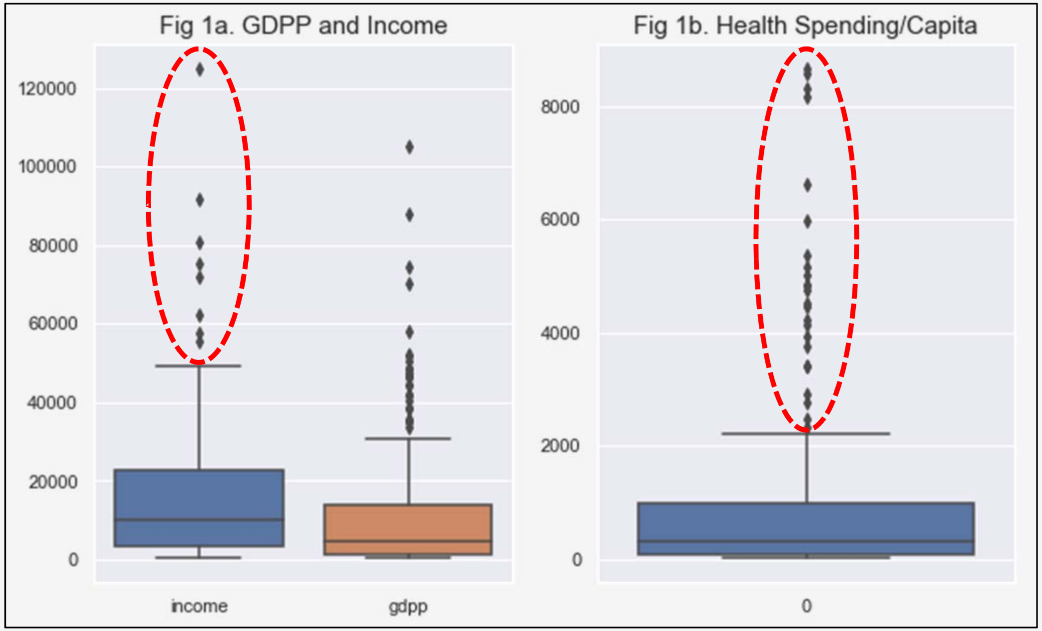
Exploring the Dataset

- **The Dataset has no missing values, thankfully.**
- **However, the dataframe.describe shows some outliers in the income and gdpp.**
- **We convert the health column into health spending per capita.**
- **Notice the max values in Income, Health and GDPP. They are clearly suffering from outliers.**

Data	columns	(total 9 Columns)	Data Type
child_mort	167	non-null	float64
exports	167	non-null	float64
health	167	non-null	float64
imports	167	non-null	float64
income	167	non-null	int64
inflation	167	non-null	float64
life_expec	167	non-null	float64
total_fer	167	non-null	float64
gdpp	167	non-null	int64

	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
count	167	167	167	167	167	167	167	167	167
mean	38.27	41.11	1056.73	46.89	17144.69	7.78	70.56	2.95	12964.16
std	40.33	27.41	1801.41	24.21	19278.07	10.57	8.89	1.51	18328.70
min	2.6	0.109	12.8212	0.0659	609	-4.21	32.1	1.15	231
25%	8.25	23.8	78.54	30.2	3355	1.81	65.3	1.80	1330
50%	19.3	35	321.89	43.3	9960	5.39	73.1	2.41	4660
75%	62.1	51.35	976.94	58.75	22800	10.75	76.8	3.88	14050
max	208	200	8663.6	174	125000	104	82.8	7.49	105000

Removing Outliers

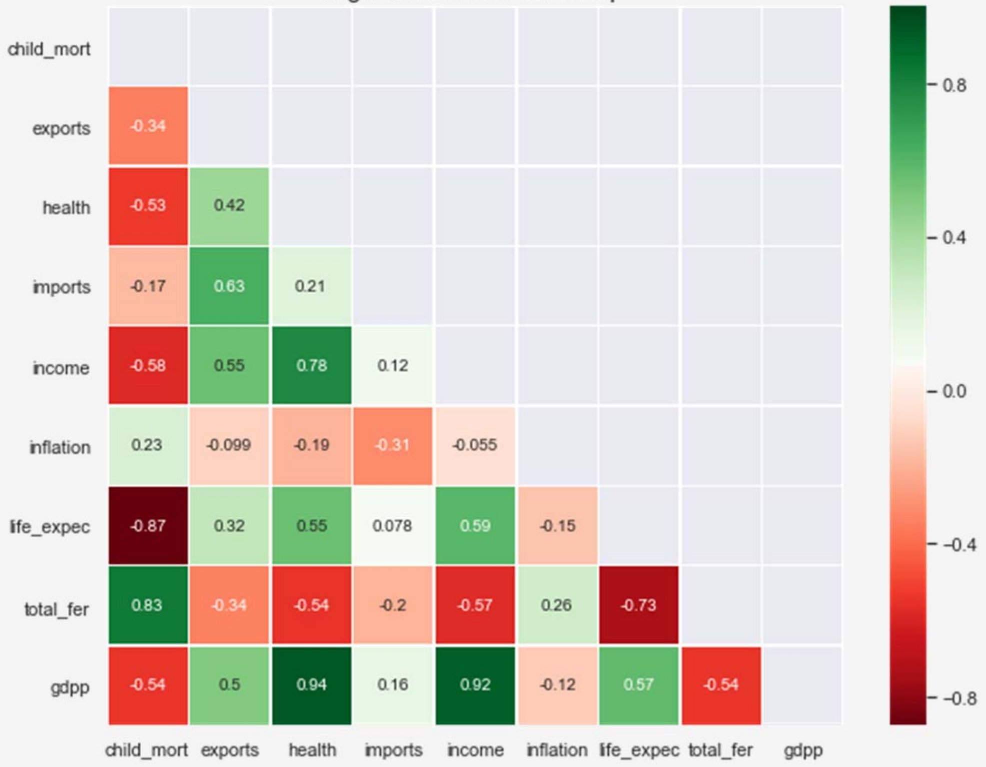


- **Our objective is to identify countries which are desperate for assistance.**
- **So , we do not need to address countries which have high income and very high spending on health.**
- **These countries would unnecessarily move the clusters in undesired manner.**
- **We eliminate the outliers in Income and Health , which are above upper whiskers.**

Scaling and Heatmap

- **We scale the data using a standard-scaler as it is the key step before PCA.**
- **A correlation heatmap shows some clear indicators of strong correlations.**
- **Child Mortality has a strong –ive correlation with GDPP, Income, Life Expectancy. This shows that abject poverty may cause high child mortality.**
- **A +ive correlation of Child Mortality with Total-Fertility also shows that countries with less emphasis on Population Control also results in high Child Mortality.**
- **On the contrary GDPP is positively correlated with Health and Life Expectancy.**
- **Notice the +ive impact on Life Expectancy due to spending on health.**

Fig 2. Correlation Heatmap



Identifying Principle Components

- **The data-set has 9 numeric features and 1 non-numeric column “Countries”.**
- **Thus our data has 9 dimensions, that we subject to PCA. Our objective is to find an alternative basis with decreasing order of explained variance.**
- **A Scree plot below shows that 6 principle components can explain 98% of explained variance.**
- **More than 50% of the variance is explained in the first two principle components itself.**

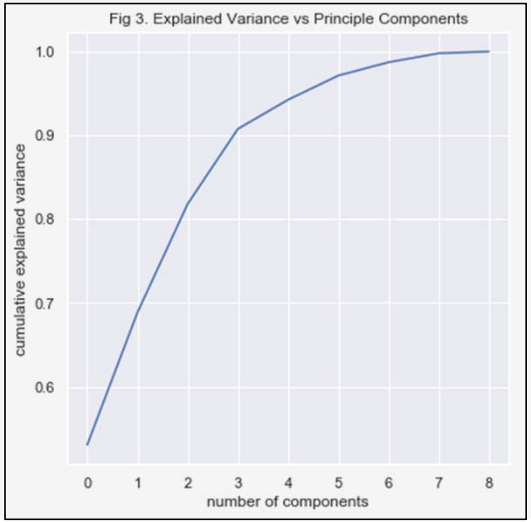


Fig 3. Explained Variance vs Principle Components

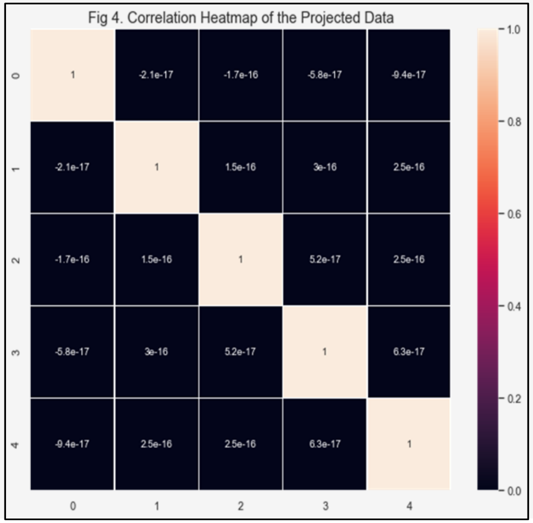


Fig 4. Correlation Heatmap of the Projected Data

	0	1	2	3	4
0	1	-2.1e-17	-1.7e-16	-5.8e-17	-9.4e-17
1	-2.1e-17	1	1.5e-16	3e-16	2.5e-16
2	-1.7e-16	1.5e-16	1	5.2e-17	2.5e-16
3	-5.8e-17	3e-16	5.2e-17	1	6.3e-17
4	-9.4e-17	2.5e-16	2.5e-16	6.3e-17	1

- **The original data is then projected onto the 6 PCs.**
- **We calculate the correlation matrix again and we notice that the off-diagonal entries are 0.**
- **This was expected , as all the new PCs are orthogonal to each other.**

K- Means Clustering (Choosing K)

- **We perform on the Clustering on the projected data on the 6 Principle Components**
- **One of the key initial steps is to identify K.**
- **We choose Elbow method and Silhouette Scores and try to figure out an optimum value of K.**
- **Silhouette doesn’t get us very encouraging results. This is because , a good K would be complemented by a S approaching 1, but it doesn’t. Nevertheless, the highest S Score is at K=4**
- **We choose K=4.**

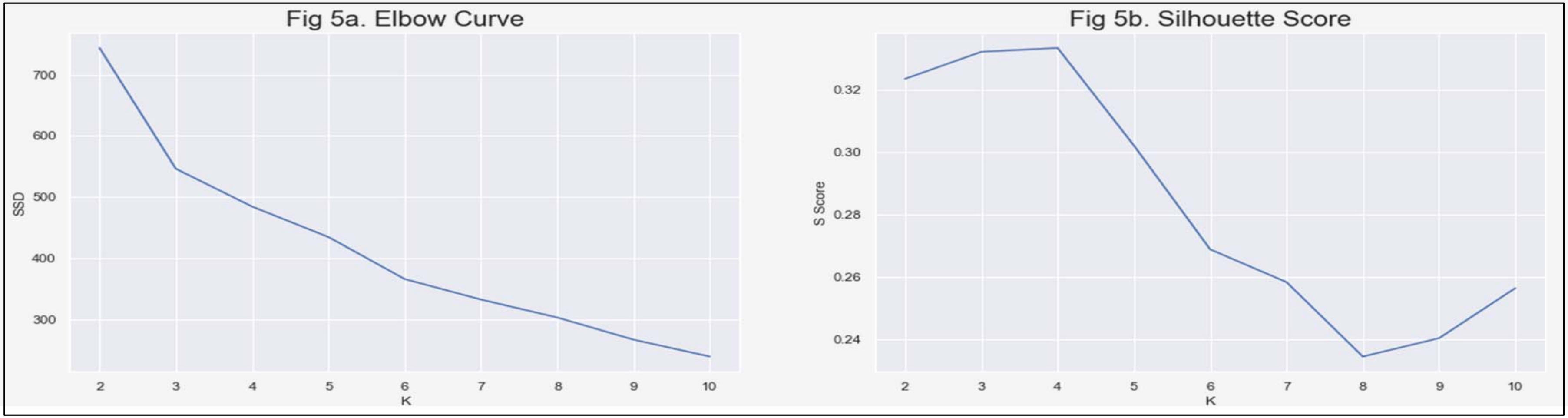


Fig 5a. Elbow Curve

Fig 5b. Silhouette Score

K- Means Results (Box Plots)

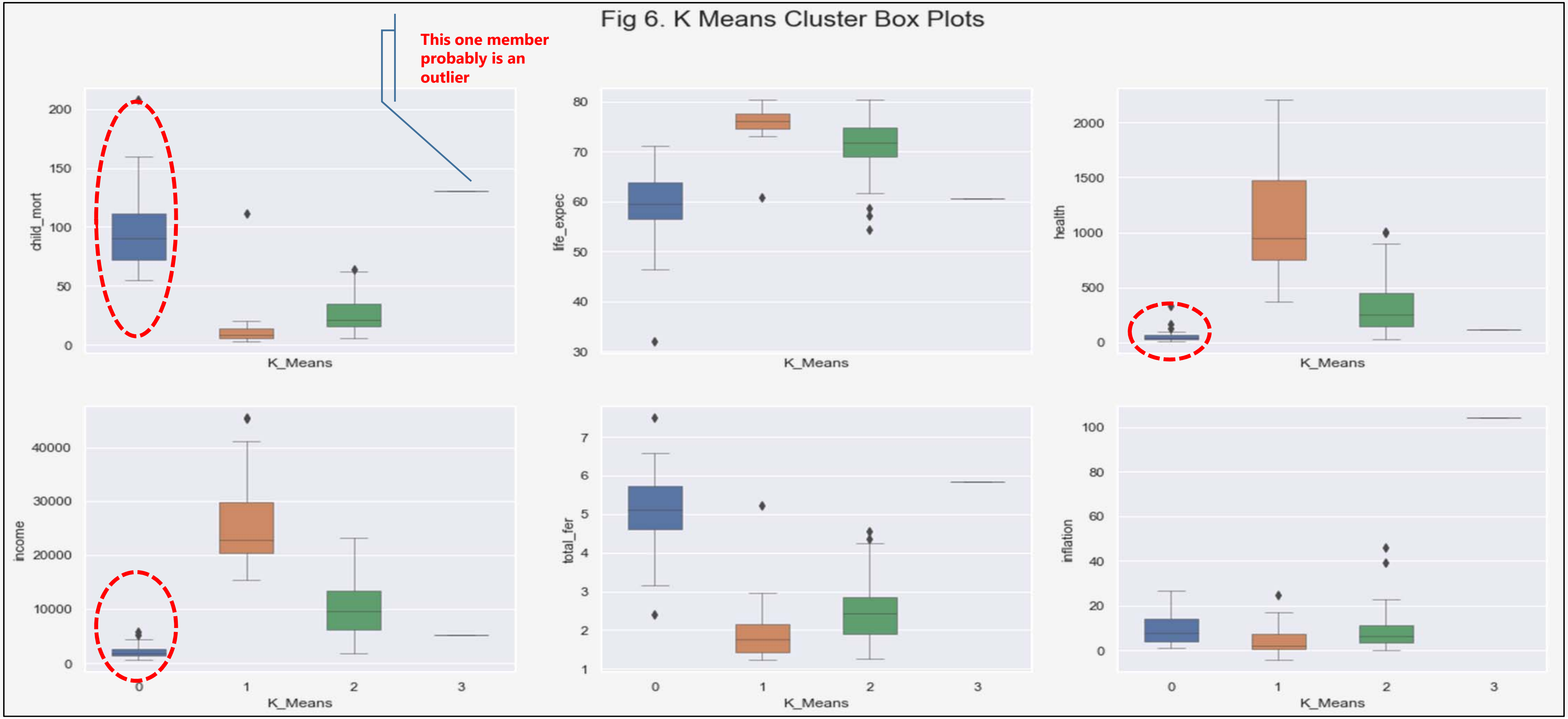


Fig 6. K Means Cluster Box Plots

This one member probably is an outlier

K- Means Results (Box Plots Cont'd) and Visualization

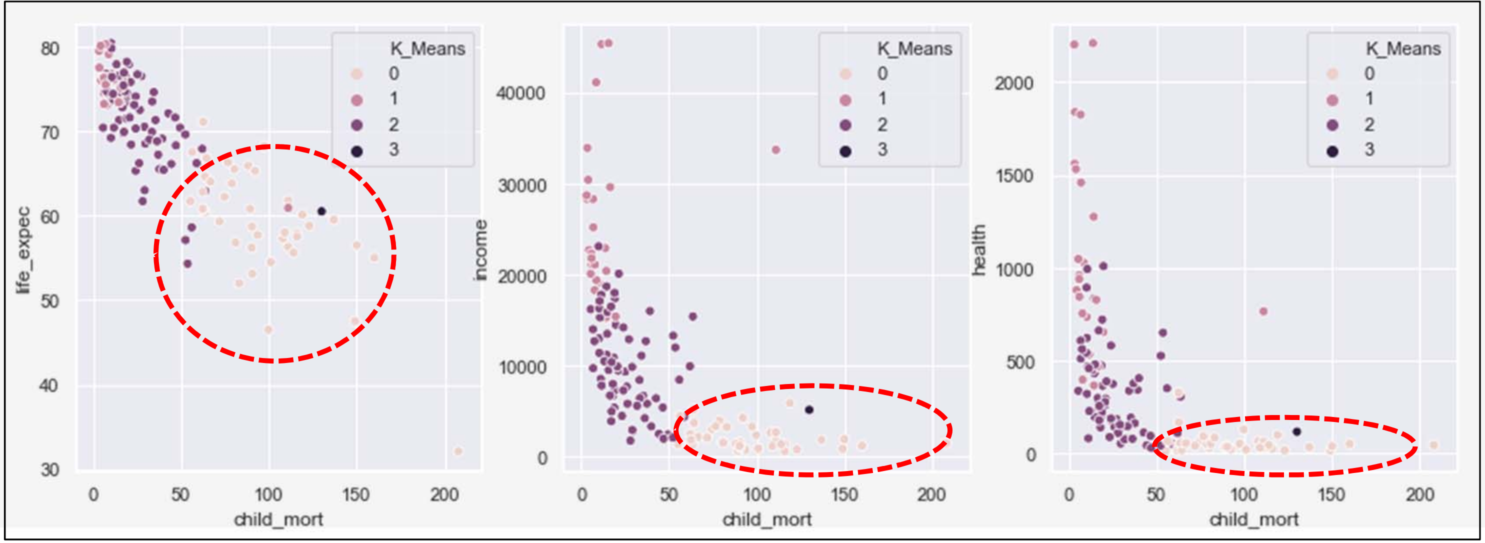
- **The strongest economies (i.e Highest income and GDP/C) are also the ones with highest mean spending on Health.**

- **These countries also have the lowest Child Mortality.**

- **On the other hand , we also notice that Cluster 0 has the worst Life Expectancy, Income, GDP/C.**

- **Also note Total Fertility, which shows practically no control on population for countries with high Child Mortality.**

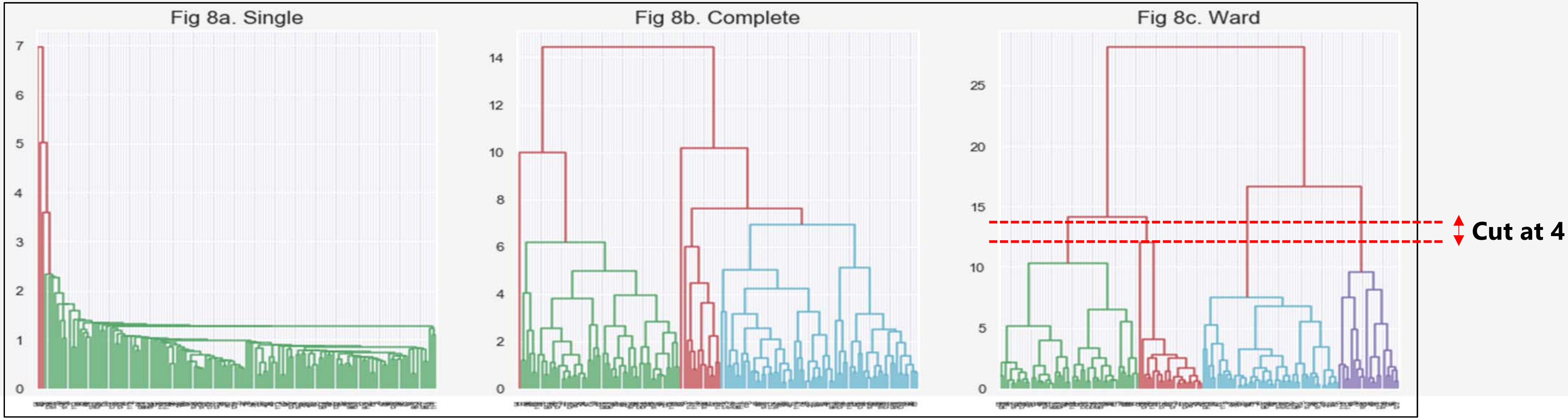
- **The pair-wise plots with themes set to Cluster Levels show that Cluster 1 needs the most attention. These are also nations , which have very low income.**



Hierarchical Clustering

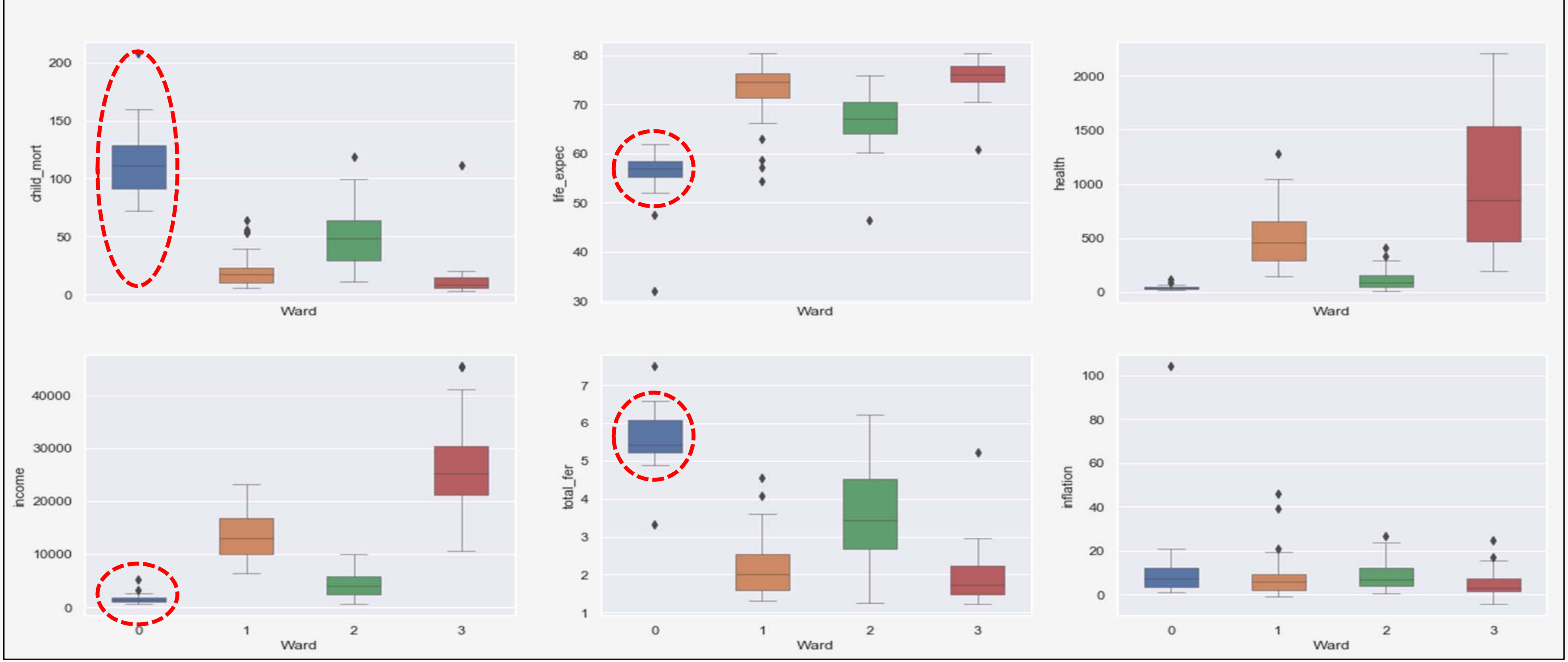
- **We try three linkage methods Single, Complete and Ward.**
- **The Single and Complete dendrograms look pretty inconclusive.**
- **The Ward method seems to produce best results.**
- **This is because from Threshold 11 to 15 , we can cover the maximum vertical distance without cutting through any other branch.**
- **We will use the labels as generated by the Ward method.**

- **The Complete Linkage produces just one Label = 3.**
- **We had a similar issue in K-Means.**
- **Hence , we select the Ward Method**



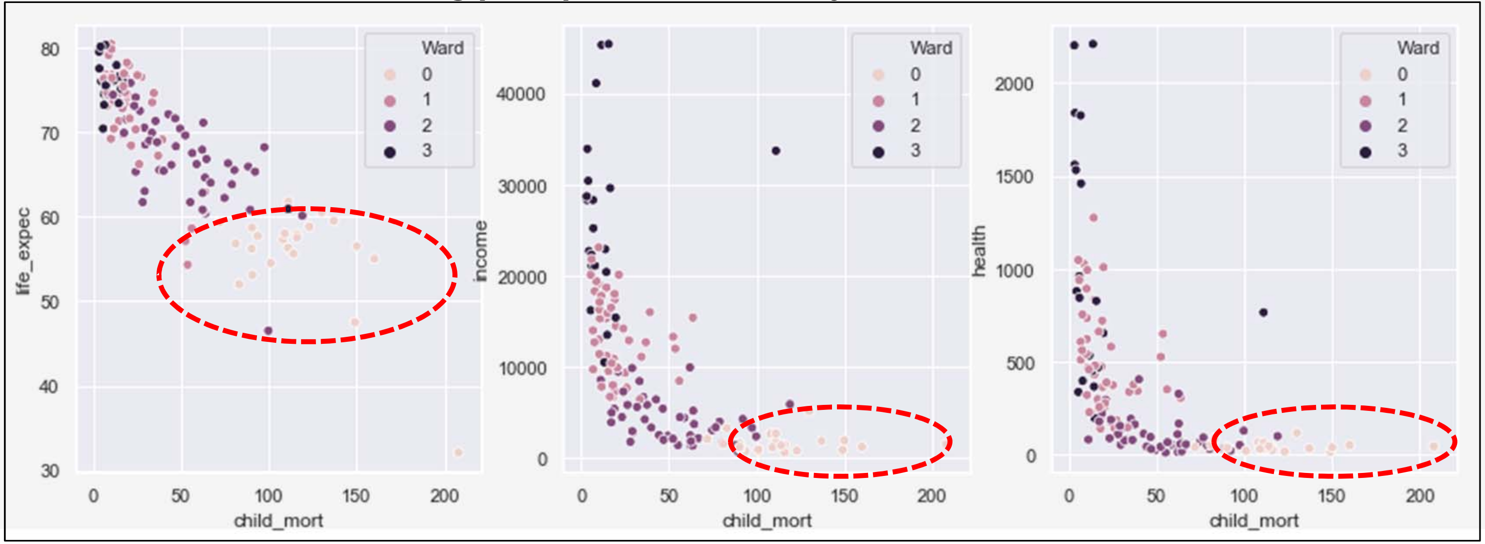
Hierarchical Results (Box Plots)

Fig 5. Hierarchical Box Plots



Hierarchical Results (Box Plots Cont'd) and Visualization

- **We come to almost the same conclusions like K-Means, however the box plots are much neatly separated.**
- **The countries with the highest Child Mortality are also the ones that have lowest income and lowest spending on health.**
- **These countries also have total-fertility very high. This shows poor presence of family planning.**
- **Interestingly , we also see a very distinct separation of Cluster 0 with high Inflation as well.**
- **On the contrary, countries with strong economies (high income and GDPP) also have very low child mortality. Their inflation is also rock bottom, indicating perhaps a stable currency.**



The Countries in Dire Need

country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	K_Means	Ward
Congo, Dem. Rep.	116	41.1	26.41	49.6	609	20.8	57.5	6.54	334	0	0
Burundi	93.6	8.92	26.79	39.2	764	12.3	57.7	6.26	231	0	0
Niger	123	22.2	17.95	49.1	814	2.55	58.8	7.49	348	0	0
Central African Republic	149	11.8	17.75	26.5	888	2.01	47.5	5.21	446	0	0
Mozambique	101	31.5	21.82	46.2	918	7.64	54.5	5.56	419	0	0

- **We use the labels from *Hierarchical Method/Ward Linkage* and attach them to the original dataset.**

- **After sorting the dataset in terms of income , we notice that worst 5 nations have been identified by both K-Means and Hierarchical Methods.**

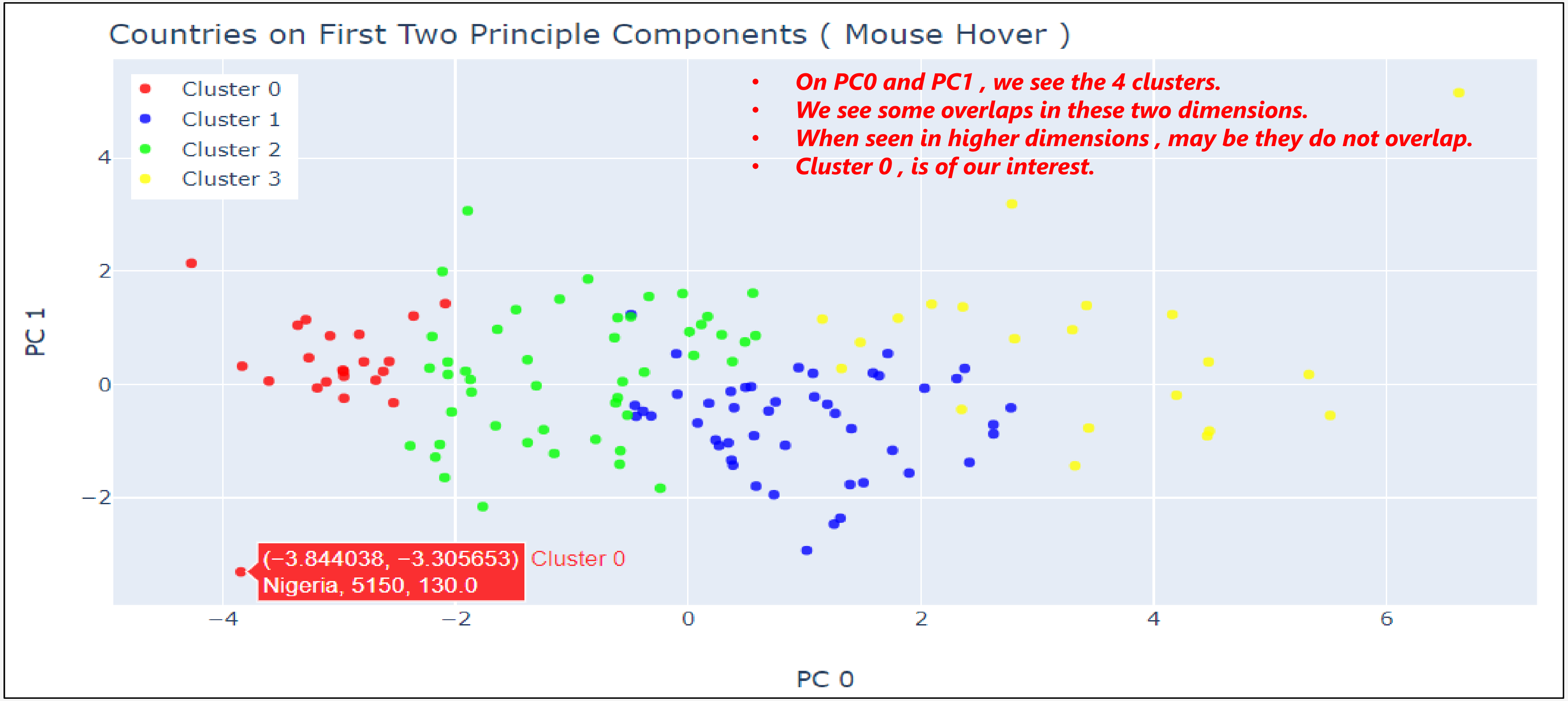
- **Their economy needs uplift , as it'd directly impact income. Improved quality of life would also affect Life Expectancy.**

- **They not only need to elevate their income , but also put a serious emphasis on Total Fertility. This requires awareness programs towards Family Planning.**

- **Spending on health is also of critical importance !!**

The Final Visualization

- **Our final visualization allows to hover the mouse and see the countries along with their Income and Child-Mortality in one plot.**



- ***On PC0 and PC1 , we see the 4 clusters.***
- ***We see some overlaps in these two dimensions.***
- ***When seen in higher dimensions , may be they do not overlap.***
- ***Cluster 0 , is of our interest.***

THANK YOU