OVERVIEW OF THE PROBLEM STATEMENT…

Our NGO, HELP International strengthens poor countries by fighting poverty and providing the citizens of these countries with basic amenities.
In a recent funding, our organization has accumulated around 10 million US Dollars, our well-intentioned CEO now has a task upon him to best use these funds…
So, we as his data analyst team are supposed to categorize the countries using some socio-economic and health factors and basically narrow it down to at least 5 countries, which are in dire need of these funds and maybe can make the most out of this.

Now, we are supposed to cluster different countries into several groups based on some socio-economic factors.

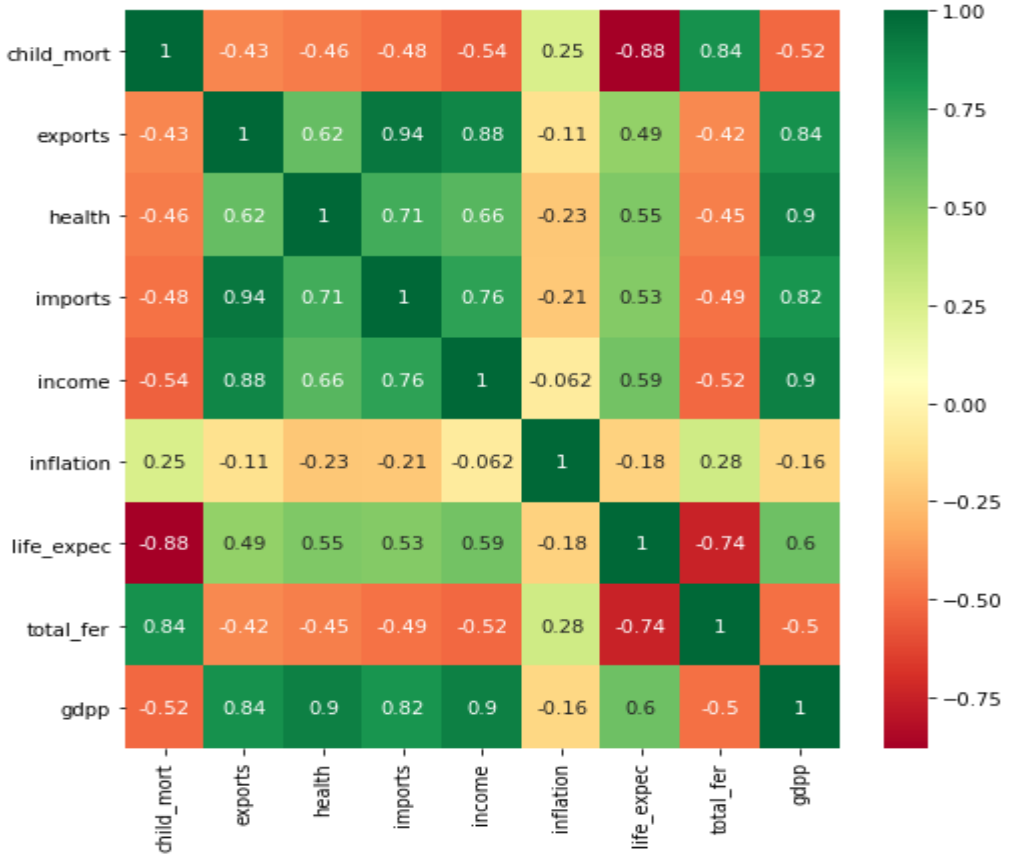
Now, we'll be starting with our analysis portion.. This PPT is aimed at making the masses understand what we did in the simplest terms possible : -

- Firstly, we started by importing the necessary libraries that are tasked at making our analysis easier.
- Next, we'll import the provided dataset to perform our analysis.
- Now, we look at the number of rows and columns that we'll be working with, and also in the process looked at some descriptive statistics like mean, median etc. of each column.
- We move onto inspecting whether there are any missing values in the data, and to our surprise there were none. So that's a nice fortune.
- Now, we convert the columns(namely, imports, exports and health) into their absolute terms from percentage terms.
- And, now we get to a little trickier section i.e. "Outlier Detection", we approached it using plain old boxplots, which clearly proved that our dataset had certain outliers.

To amend this we came up with our **own theory**, that **we'll cap stark negative indicators in the lower range like, child mortality, inflation and total fertility of an average woman, at 25 percentile**, meaning any observation below 25 percentile of these values will be removed. **This was done to remove some overly Developed nations so that they don't hinder with our grouping.**

And, for **other indicators such as income, GDP per capita, imports, exports and health we capped the upper band of observations with above 90 percentile of these values.**
So as to not remove meaningful observations the 90 percentile capping limit was chosen.
And , due to these specific constraints <u>we got rid of 17 overly developed nations that could overpower their underdeveloped peers.</u>

- We used several visualisations at our disposal to inspect the distribution of different features, and their relationship with other features using pair plots and a correlation matrix



This right here is what a correlation matrix looks like, its main purpose is to indicate in form of a decimal number, in which way one column influences the other column.
Like, Income here is greatly influenced in a positive(green) sense by exports and gdpp columns.

Now, we move onto clustering using the K-Means Algorithm..

- Firstly, we check using Hopkin's statistic, whether our data is cluster able or not, a score greater than 0.5 indicates good cluster tendency. We got a score of around 0.9, which is great.

(Hopkin's statistic basically compares our data with a random uniformly distributed data and measures the dissimilarity between our data and the randomly generated data.)

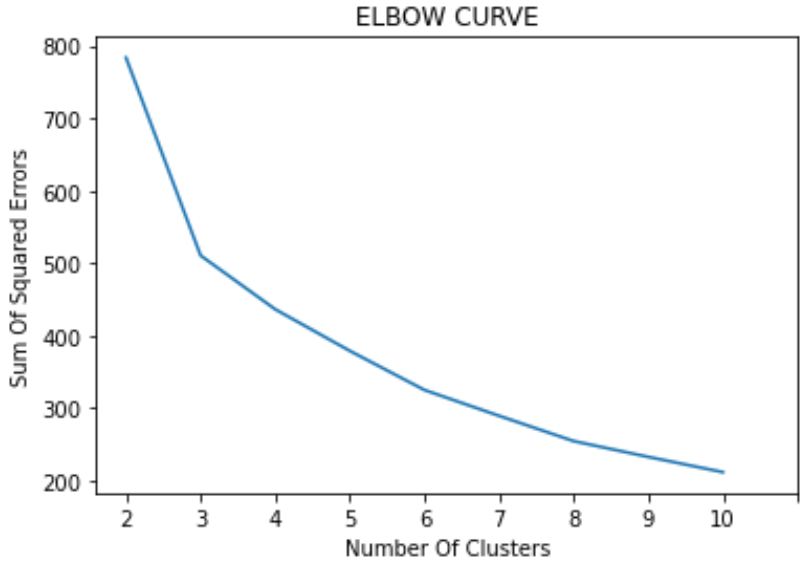
- Then, we move onto standardising our data, so that each column has a mean of 0 and a standard deviation of 1, which can be confusing.

In a layman's terms , we brought all our columns to the same scale so that the absolute difference between them does not trouble us.

We, now, have to determine the "K" in K-means i.e. the number of clusters that we want.

- We employ elbow curve that measures the average sum of squared differences between cluster centres and our data points.

Basically, we choose the number of cluster which drastically reduces our sum of squared differences. That magical number for us comes to be "3" clusters as is visible in the line chart on the next slide.

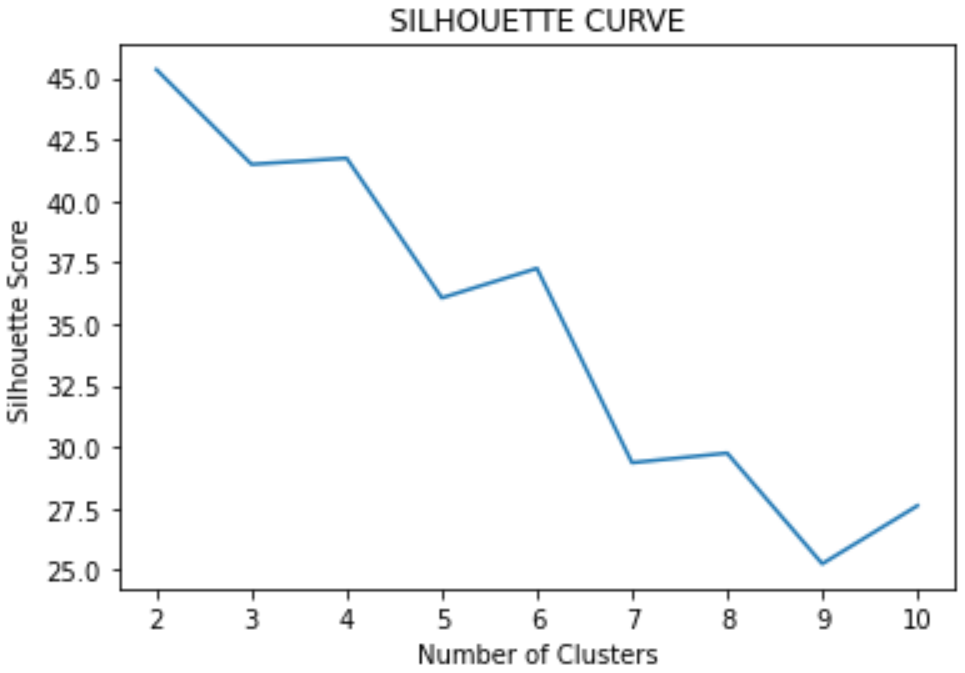


The below curve is the "Silhouette Curve", which shows how many clusters will best suit our data.
Silhouette score maximises the inter-cluster distance, while minimizing the intra-cluster distance. The larger the score, the better it is.
Now, we **chose 3 as the number of clusters** that we'll be working with, we came up with this by combining the results from both these curves.

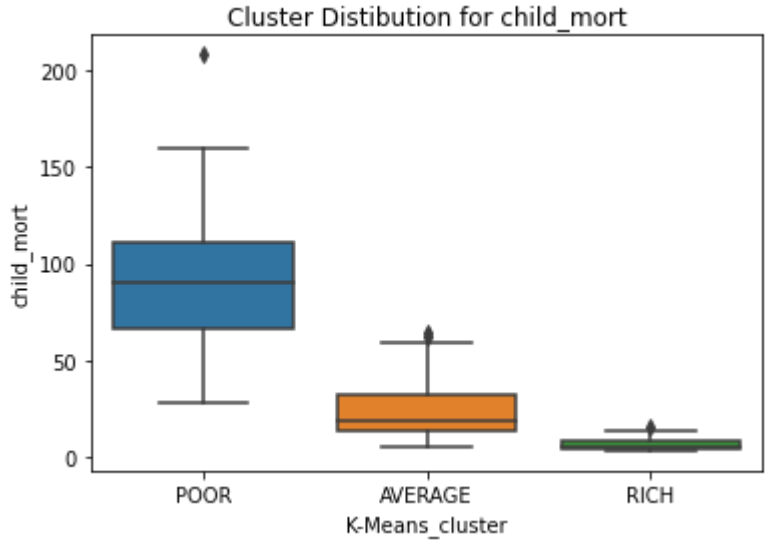
<u>Some might ask, why didn't we go with 4 clusters as it yields a little better Silhouette score as is evident in the below snip ?</u>

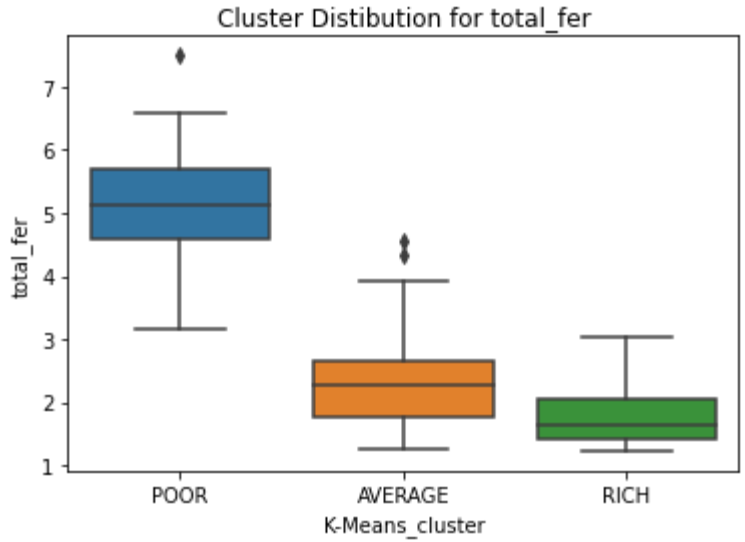
When we dig deep we find that "Nigeria" is the reason for this phenomenon, which performs above average on Economic grounds, but way below on the Health front, hence there is no sense in making a whole other cluster just for 1 participant, hence Nigeria will be grouped with other poor countries in 3 clusters.

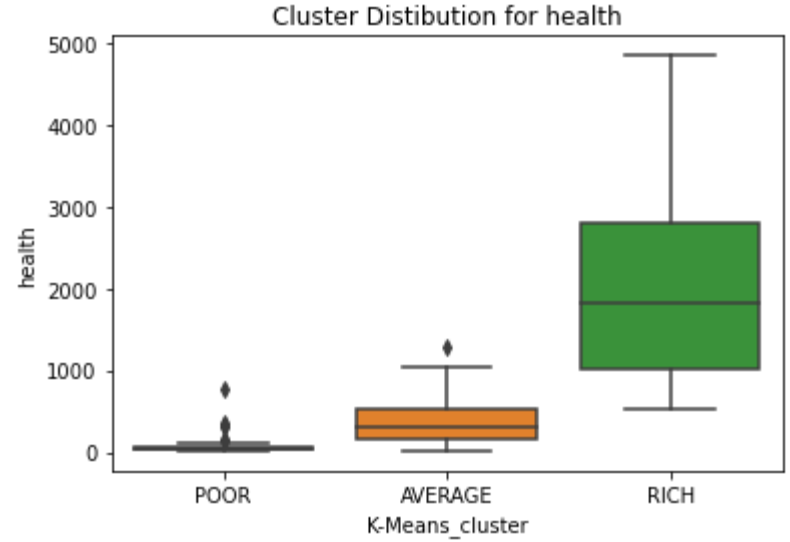
```
The Silhouette score for 2 clusters is, 45.34
The Silhouette score for 3 clusters is, 41.49
The Silhouette score for 4 clusters is, 41.74
```



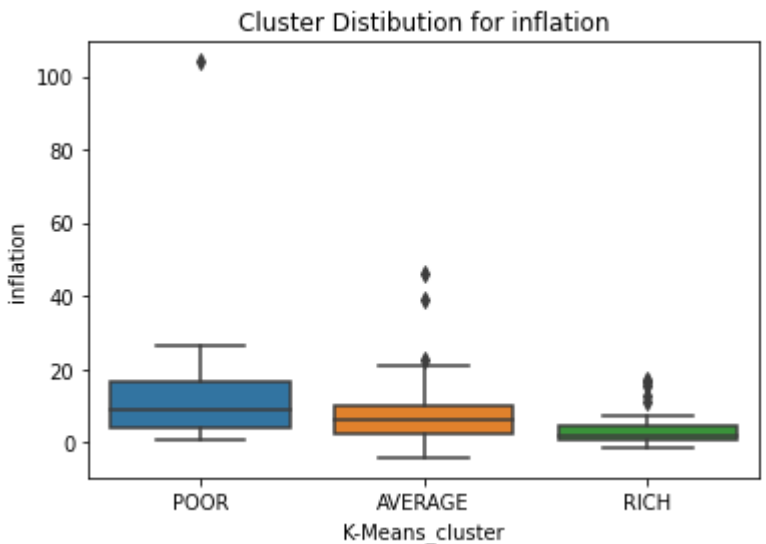
BOXPLOTS AS PER K-MEANS CLUSTER ON HEALTH INDICATORS

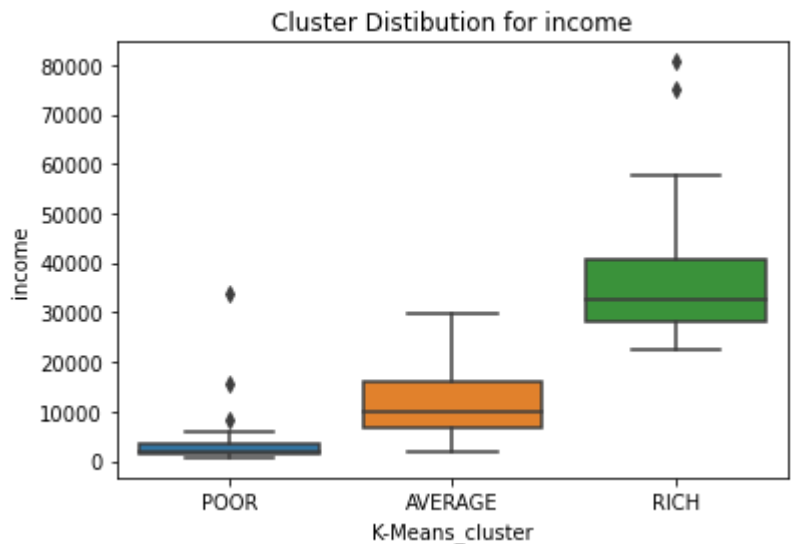


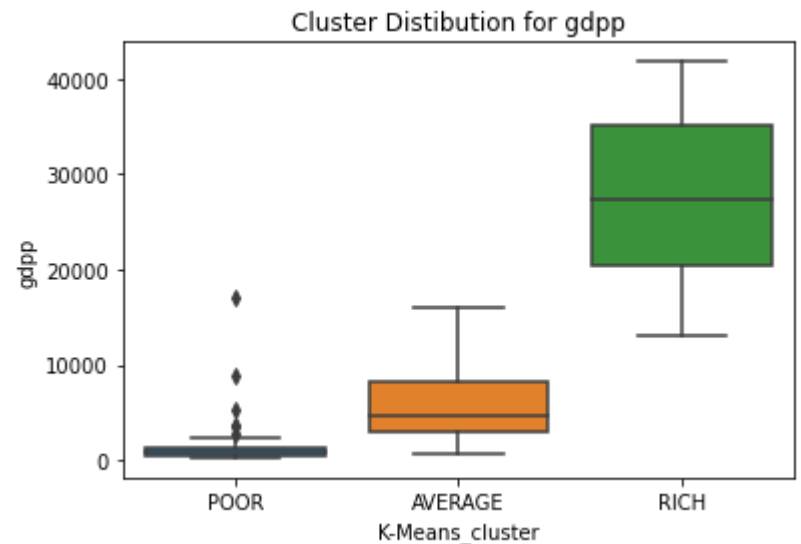


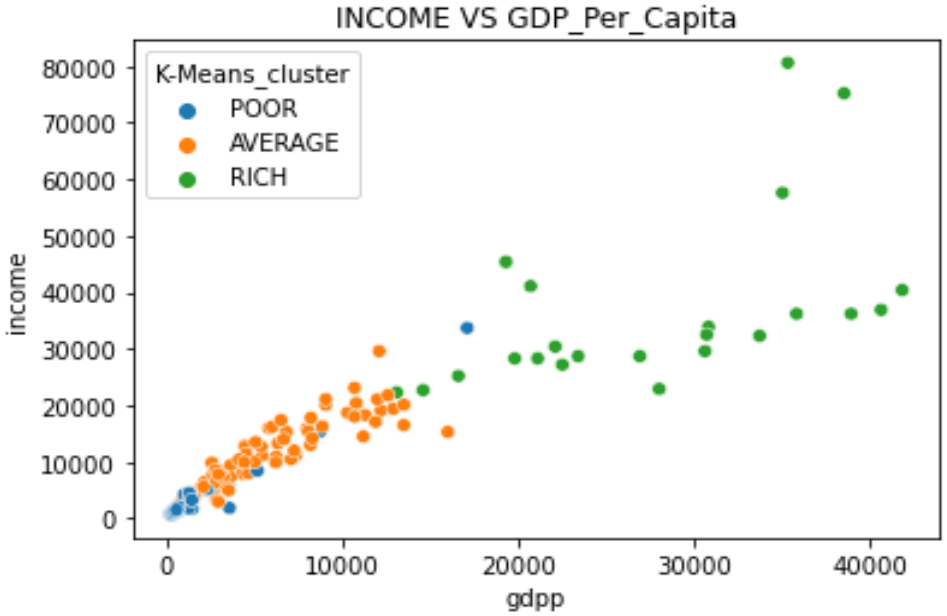


BOXPLOTS AS PER K-MEANS CLUSTER ON ECONOMIC GROUNDS



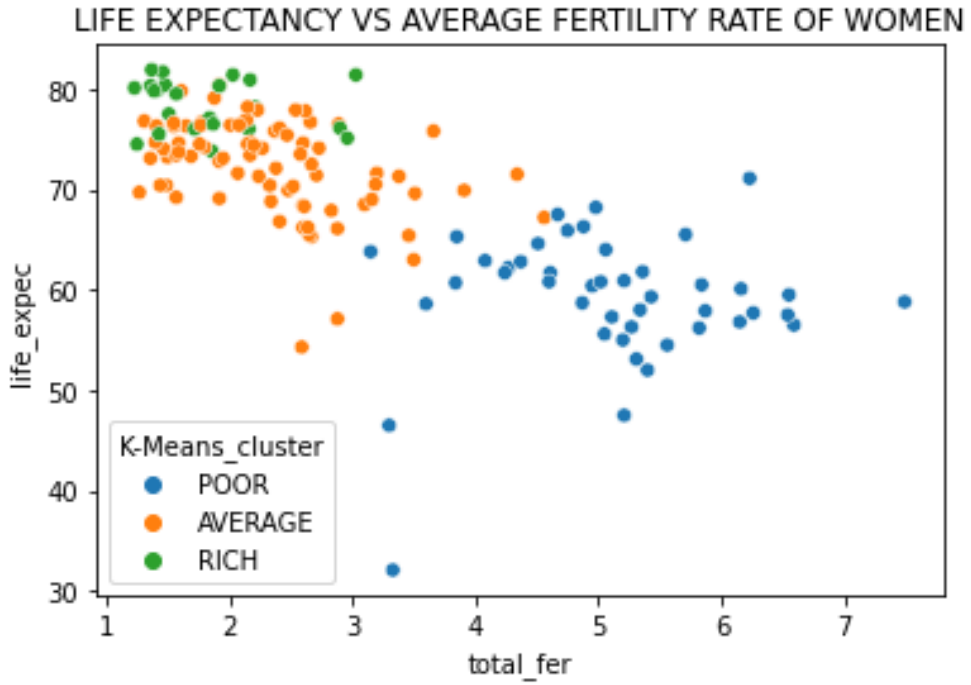


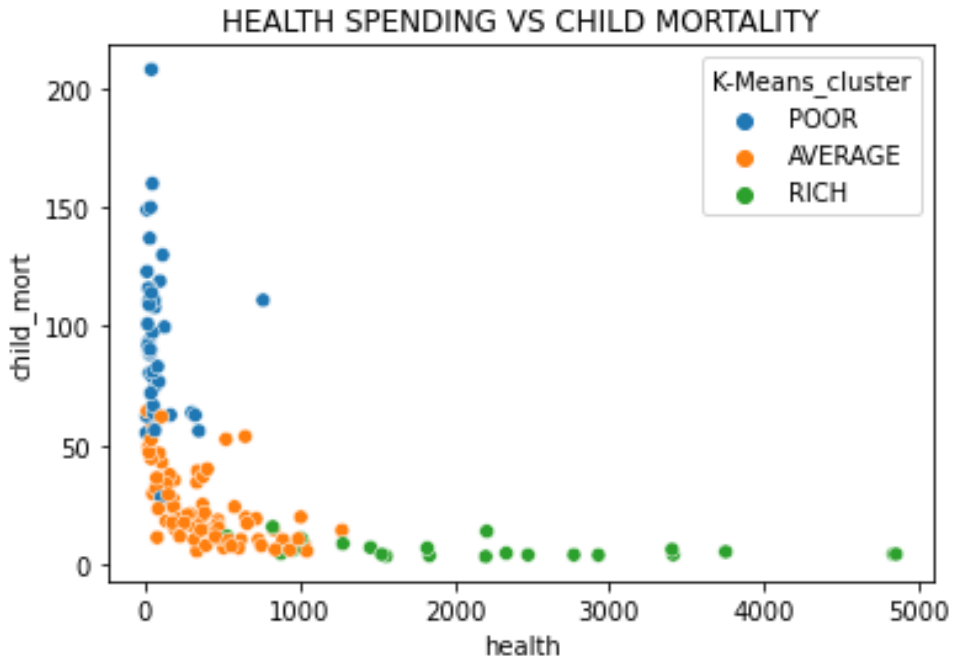




The Scatterplot INCOME v/s GDP per Capita , shows how Poor(Blue) countries have lesser GDP per capita and low income, while the Rich(Green) have higher GDP per capita and income levels..

The Scatterplot, LIFE EXPECTANCY v/s AVERAGE FERTILITY , shows how Rich(Green) have higher life expectancy and lower total fertility rates, on the other hand, Poor(Blue) have lower life expectancy and higher total fertility rates





Above Scatterplot, HEALTH SPENDING v/s CHILD MORTALITY , shows how Rich(Green) have higher health spending and lower child mortality rates, on the other hand, Poor(Blue) have lower health spending and higher child mortality rates...

These all scatterplots, bar plots and boxplots are aimed at proving that our K-Means algorithm has performed splendidly and have successfully grouped all countries into their respective clusters.

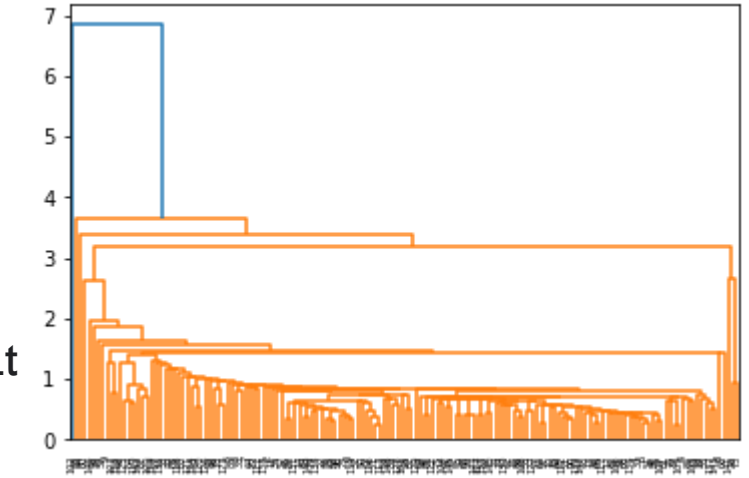
COUNTRIES THAT ARE PRIME CANDIDATES AS PER K-MEANS CLUSTERING ALGORITHM

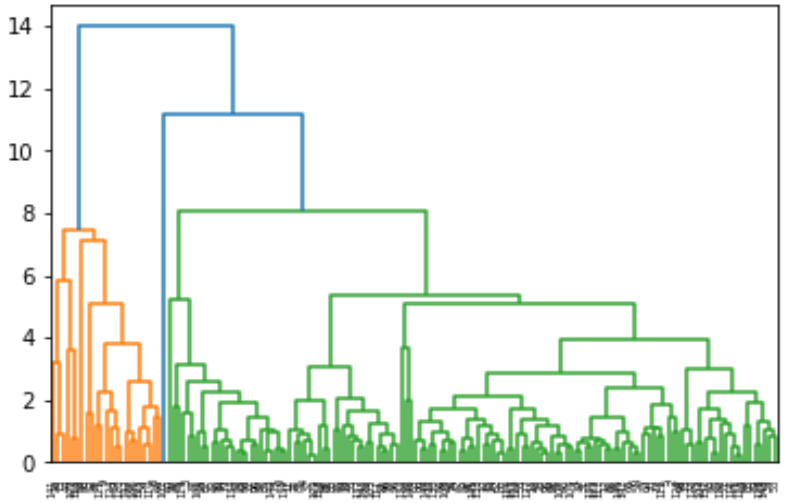
	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	K-Means_cluster
26	Burundi	93.6	20.6052	26.7960	90.552	764	12.30	57.7	6.26	231	POOR
88	Liberia	89.3	62.4570	38.5860	302.802	700	5.47	60.8	5.02	327	POOR
37	Congo, Dem. Rep.	116.0	137.2740	26.4194	165.664	609	20.80	57.5	6.54	334	POOR
112	Niger	123.0	77.2560	17.9568	170.868	814	2.55	58.8	7.49	348	POOR
132	Sierra Leone	160.0	67.0320	52.2690	137.655	1220	17.20	55.0	5.20	399	POOR
93	Madagascar	62.2	103.2500	15.5701	177.590	1390	8.79	60.8	4.60	413	POOR
106	Mozambique	101.0	131.9850	21.8299	193.578	918	7.64	54.5	5.56	419	POOR
31	Central African Republic	149.0	52.6280	17.7508	118.190	888	2.01	47.5	5.21	446	POOR
94	Malawi	90.5	104.6520	30.2481	160.191	1030	12.10	53.1	5.31	459	POOR
50	Eritrea	55.2	23.0878	12.8212	112.306	1420	11.60	61.7	4.61	482	POOR

Moving onto Hierarchical Clustering

For to use Hierarchical clustering, we have to define the type of linkage that we'll be using..

- Firstly, we went with single linkage, which focusses on minimum distances or nearest neighbor between clusters, which produced results that weren't easily interpretable

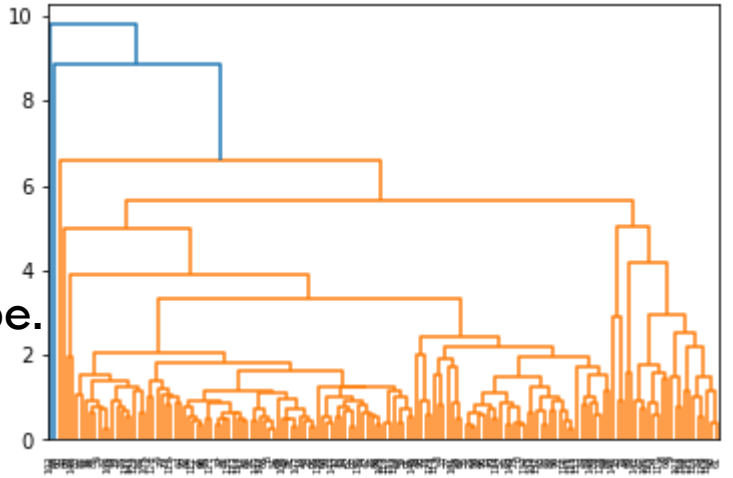




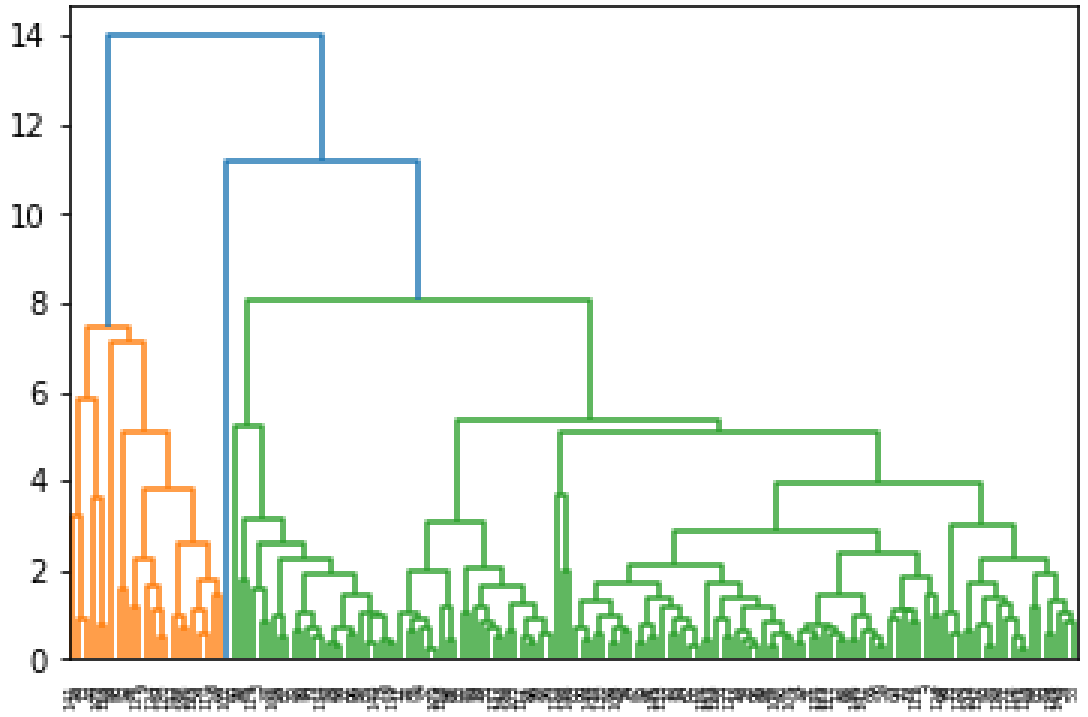
- Then, we moved onto complete linkage wherein we concentrate on maximum distance or furthest neighbor between clusters.

This time the results were pleasing to the eyes and easily interpretable

- And lastly, Just for fun, tried the average linkage type which averages the distance between clusters, but the same problem carried forward from single linkage type. The results weren't interpretable



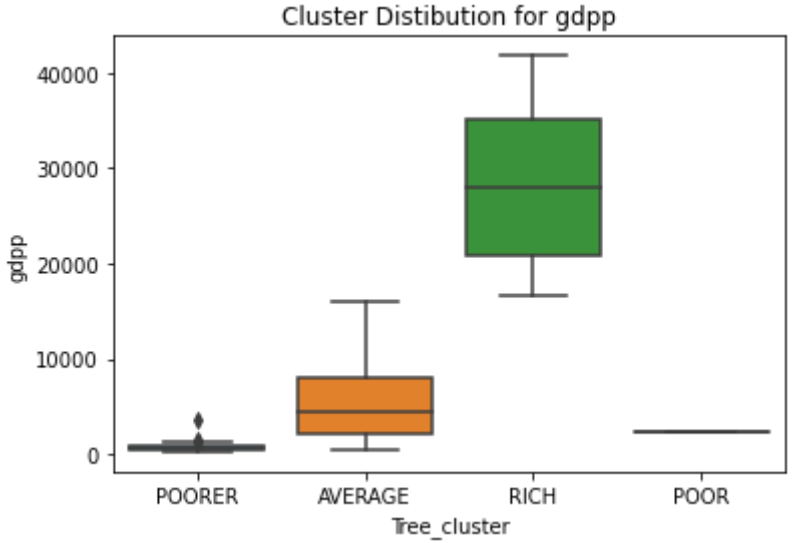
Now, the same problem occurred while choosing the number of clusters a.k.a. “Nigeria”



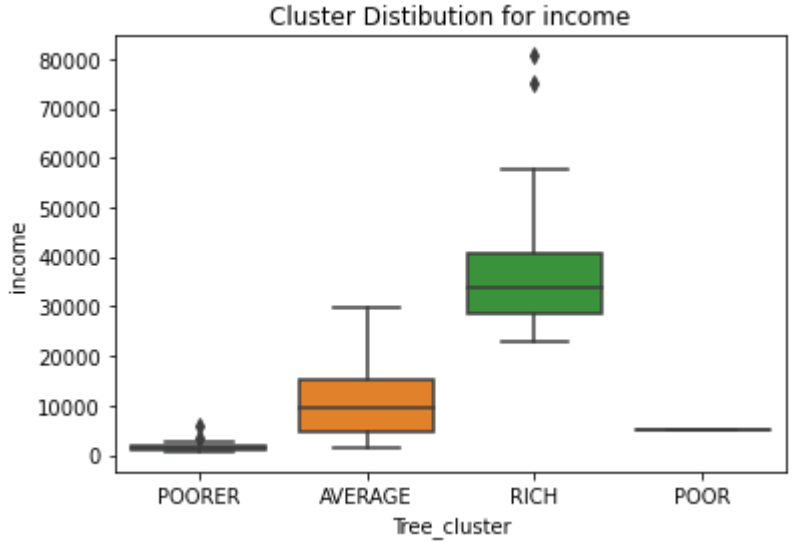
Nigeria, here, is the Blue line carried down to the x-axis, no matter what we did, Nigeria was to be a separate cluster.
And, hence, this made us classify our results to **4 clusters.**
Therefore, we cut our dendrogram at Distance(Y) = 7.5(approx.)

BOXPLOTS OF INDICATORS NAMELY CHILD MORTALITY, GDP Per Capita and INCOME AS PER HIERARCHICAL CLUSTERING ALGORITHM

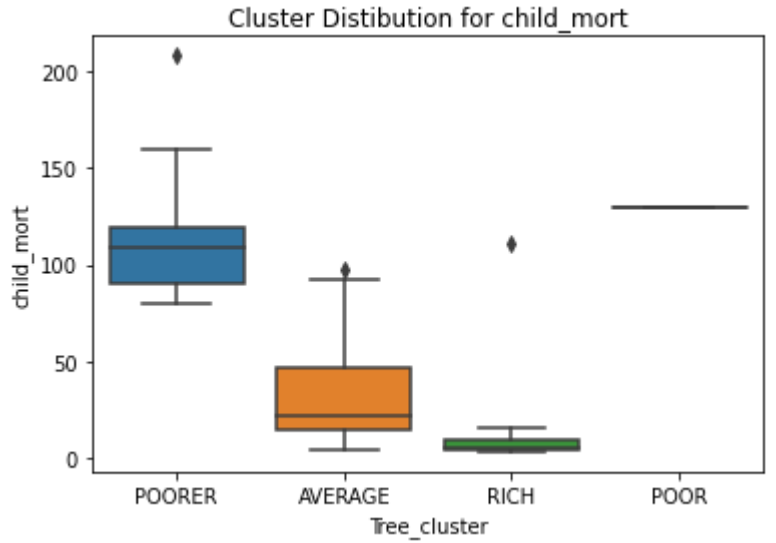
The Boxplots are reminiscent of the previous ones, except for the addition of a 4th cluster



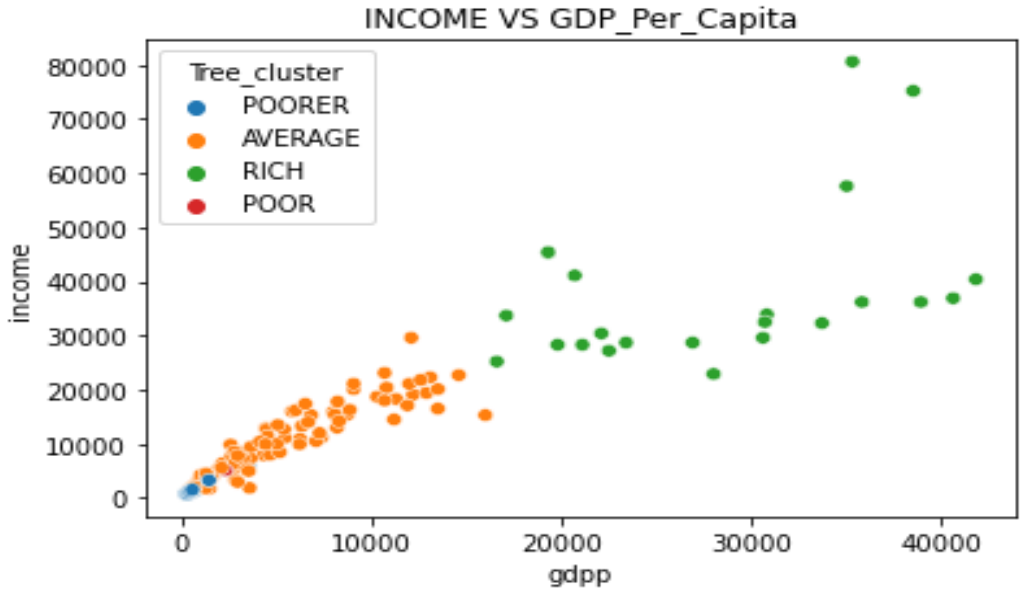
The above chart shows how Poor countries have Lower GDP per capita levels, than their Richer peers



The above chart shows how Poor countries have Lower income levels, than their Richer peers

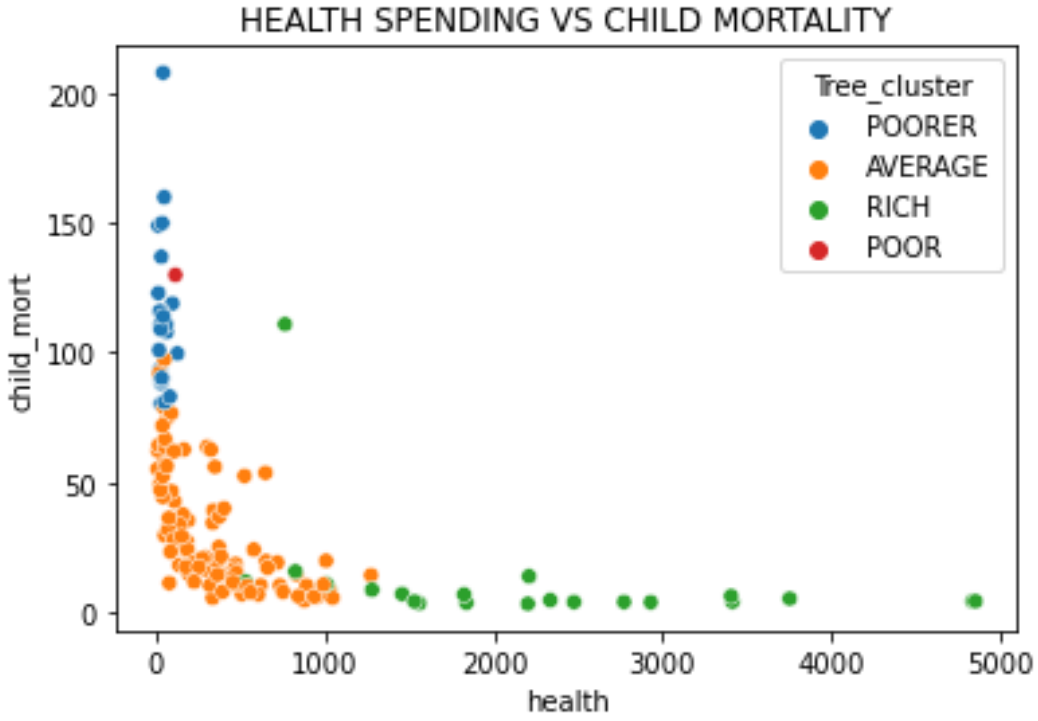


The above chart shows how Poor countries have Higher Child mortality rates, than their Richer peers



The Scatterplot INCOME v/s GDP per Capita , shows how Poor(Blue and Red) countries have lesser GDP per capita and low income, while the Rich(Green) have higher GDP per capita and income levels..

This Scatterplot, HEALTH SPENDING v/s CHILD MORTALITY , shows how Rich(Green) have higher health spending and lower child mortality rates, on the other hand, Poor(Blue and Red) have lower health spending and higher child mortality rates...



COUNTRIES THAT ARE TOP CONTENDERS AS PER HIERARCHICAL CLUSTERING ALGORITHM

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	K-Means_cluster	Tree_cluster
26	Burundi	93.6	20.6052	26.7960	90.552	764	12.30	57.7	6.26	231	POOR	POORER
88	Liberia	89.3	62.4570	38.5860	302.802	700	5.47	60.8	5.02	327	POOR	POORER
37	Congo, Dem. Rep.	116.0	137.2740	26.4194	165.664	609	20.80	57.5	6.54	334	POOR	POORER
112	Niger	123.0	77.2560	17.9568	170.868	814	2.55	58.8	7.49	348	POOR	POORER
132	Sierra Leone	160.0	67.0320	52.2690	137.655	1220	17.20	55.0	5.20	399	POOR	POORER
106	Mozambique	101.0	131.9850	21.8299	193.578	918	7.64	54.5	5.56	419	POOR	POORER
31	Central African Republic	149.0	52.6280	17.7508	118.190	888	2.01	47.5	5.21	446	POOR	POORER
94	Malawi	90.5	104.6520	30.2481	160.191	1030	12.10	53.1	5.31	459	POOR	POORER
150	Togo	90.3	196.1760	37.3320	279.624	1210	1.18	58.7	4.87	488	POOR	POORER
64	Guinea-Bissau	114.0	81.5030	46.4950	192.544	1390	2.97	55.6	5.05	547	POOR	POORER

THE ANALYST'S OPINION

Both the algorithms have performed neck-to-neck, when results are compared and the final Top10 list is generated.

But, with K-Means we had the flexibility to not let "Nigeria" become a separate cluster, while this privilege was absent in Hierarchical.

OUR FINAL 5 COUNTRIES THAT REQUIRE AID AT THE EARLIEST ARE

I. **BURUNDI**

II. **LIBERIA**

III. **DEMOCRATIC REPUBLIC OF CONGO**

IV. **NIGER**

V. **SIERRA LEONE**

These countries had the lowest GDP per Capita, highest child mortality rates and lowest incomes, as per the algorithms.