Clustering and PCA Assignment

Jaidip Ghosh

Problem Statement

- **HELP** Organisation **CEO** of the **NGO**, needs to decide how to use the money **strategically** and **effectively.**
- This can be done by choosing the countries that are in the **direst need of aid.**
- Main objective is to **categorise** the **countries** using **socio-economic** and **health factors** that will determine the overall development of the country.
- **Suggestion** will be given to **CEO** what needs to be **focus** on **most**.

Data Exploration

- Dimension of the Dataset **Country-data.csv** is **167** rows and **10** columns
- Out of **10** columns only one columns **country** is **categorical** and **rest all** are **numerical**
- **None** of the **columns** and **rows** had **Null Values**
- **No Duplicate** records were there in the dataset
- **Statistics** of the Data set, as we can see below there are Features which has the a **huge Data Spread**

	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
count	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000
mean	38.270060	41.108976	6.815689	46.890215	17144.688623	7.781832	70.555689	2.947964	12964.155689
std	40.328931	27.412010	2.746837	24.209589	19278.067698	10.570704	8.893172	1.513848	18328.704809
min	2.600000	0.109000	1.810000	0.065900	609.000000	-4.210000	32.100000	1.150000	231.000000
25%	8.250000	23.800000	4.920000	30.200000	3355.000000	1.810000	65.300000	1.795000	1330.000000
50%	19.300000	35.000000	6.320000	43.300000	9960.000000	5.390000	73.100000	2.410000	4660.000000
75%	62.100000	51.350000	8.600000	58.750000	22800.000000	10.750000	76.800000	3.880000	14050.000000
max	208.000000	200.000000	17.900000	174.000000	125000.000000	104.000000	82.800000	7.490000	105000.000000

Exploratory Data Analysis

- All the **countries** were **unique** in the dataset
- **Bivariate Analysis:** Here the Features (Numerical) were being compared with Country (Categorical)

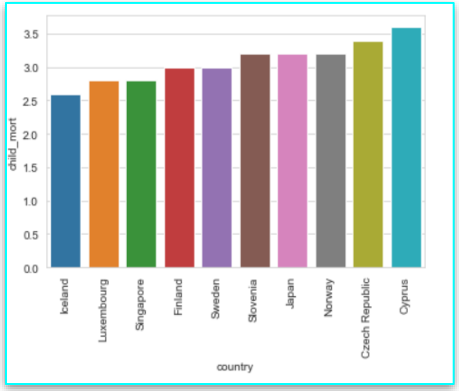


Figure 1

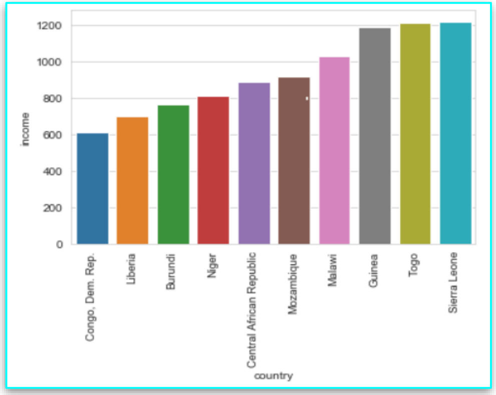


Figure 2

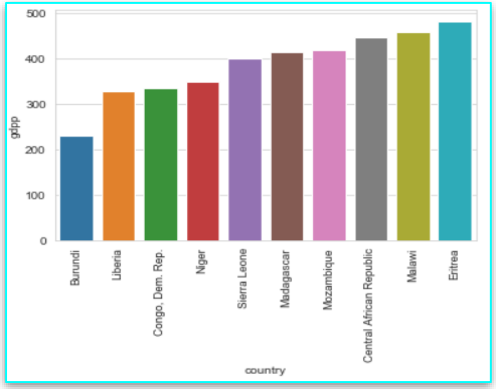


Figure 3

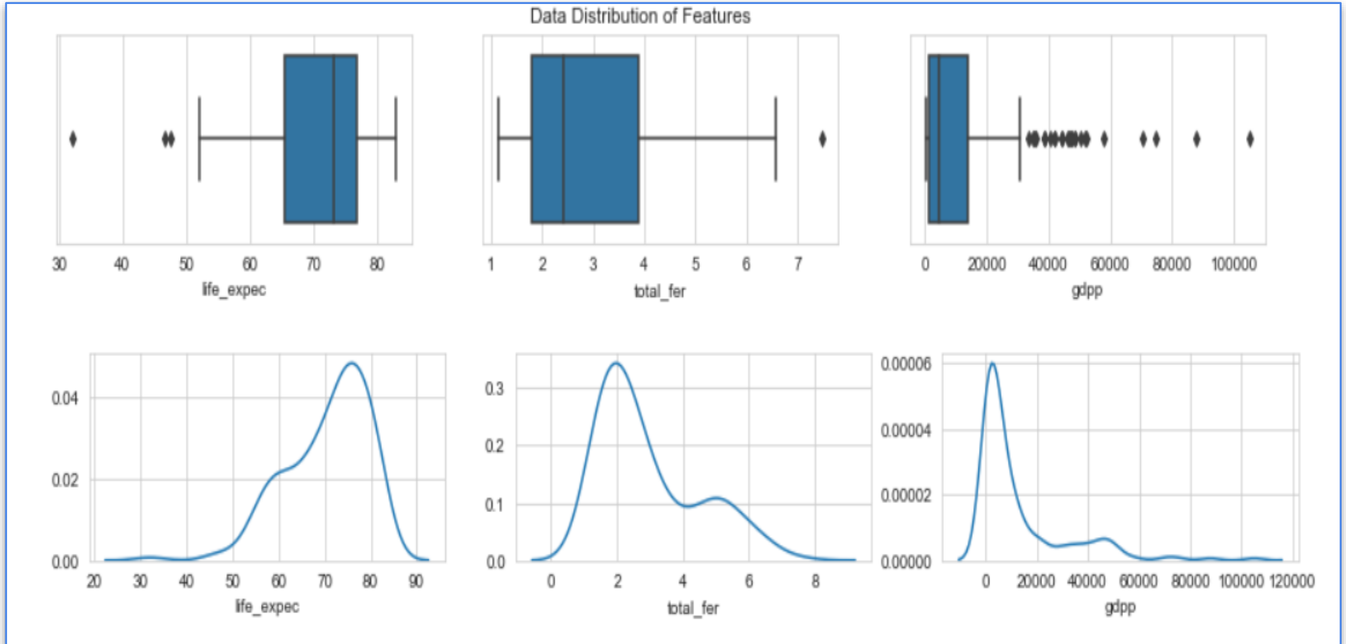
This plots shows the distribution for bottom 10 country for these features

Exploratory Data Analysis

- **Univariate Analysis:** Analysis about the Data Spread of each Features

For all of the **Features** the **Data Spread** were seen, so as to visualize the **distribution pattern**.

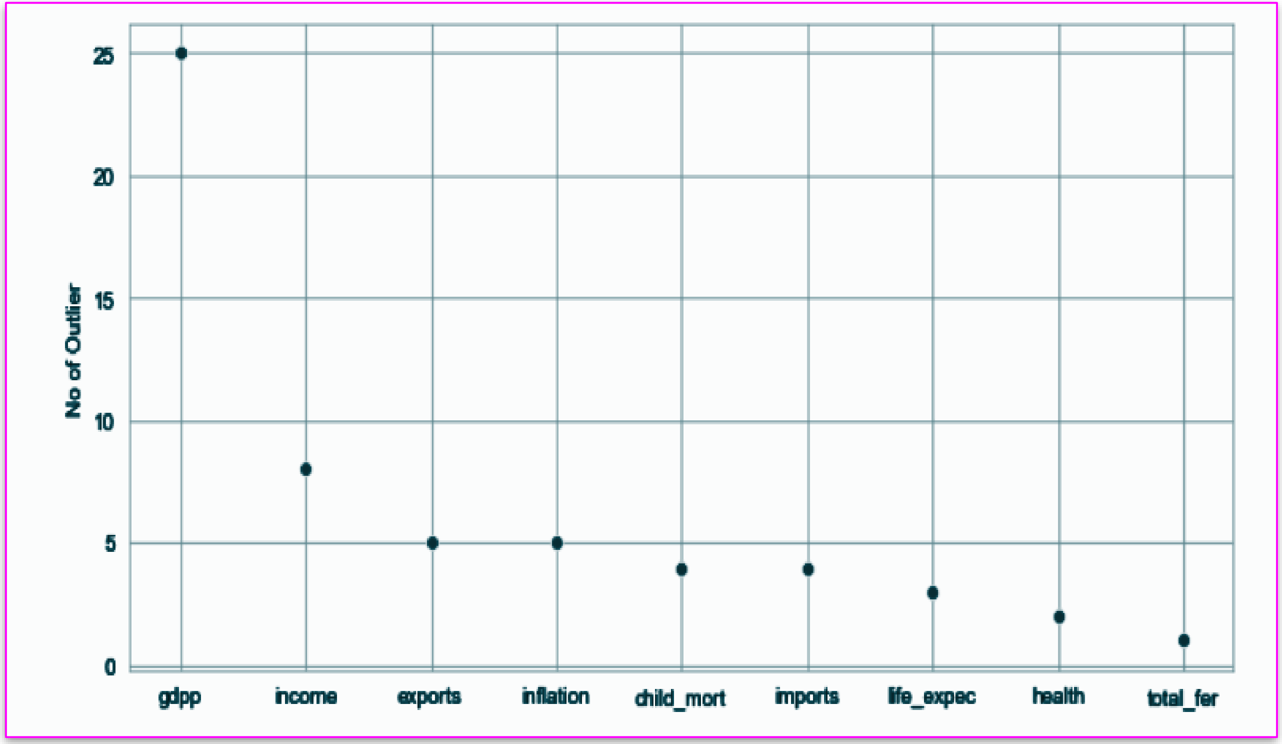
There we found that there are some points which lies beyond the **Upper and Lower whisker** and also the skewness was checked




```
Skewness for life_expec is -0.97
Skewness for total_fer is 0.97
Skewness for gdpp is 2.22
```

Exploratory Data Analysis

Outlier Analysis



How was it plotted ?

- Found the Q1 and Q3 and then found the **Interquartile Range** (IQR) method, the data point which were less than **Q1 - (1.5 * IQR)** and more than

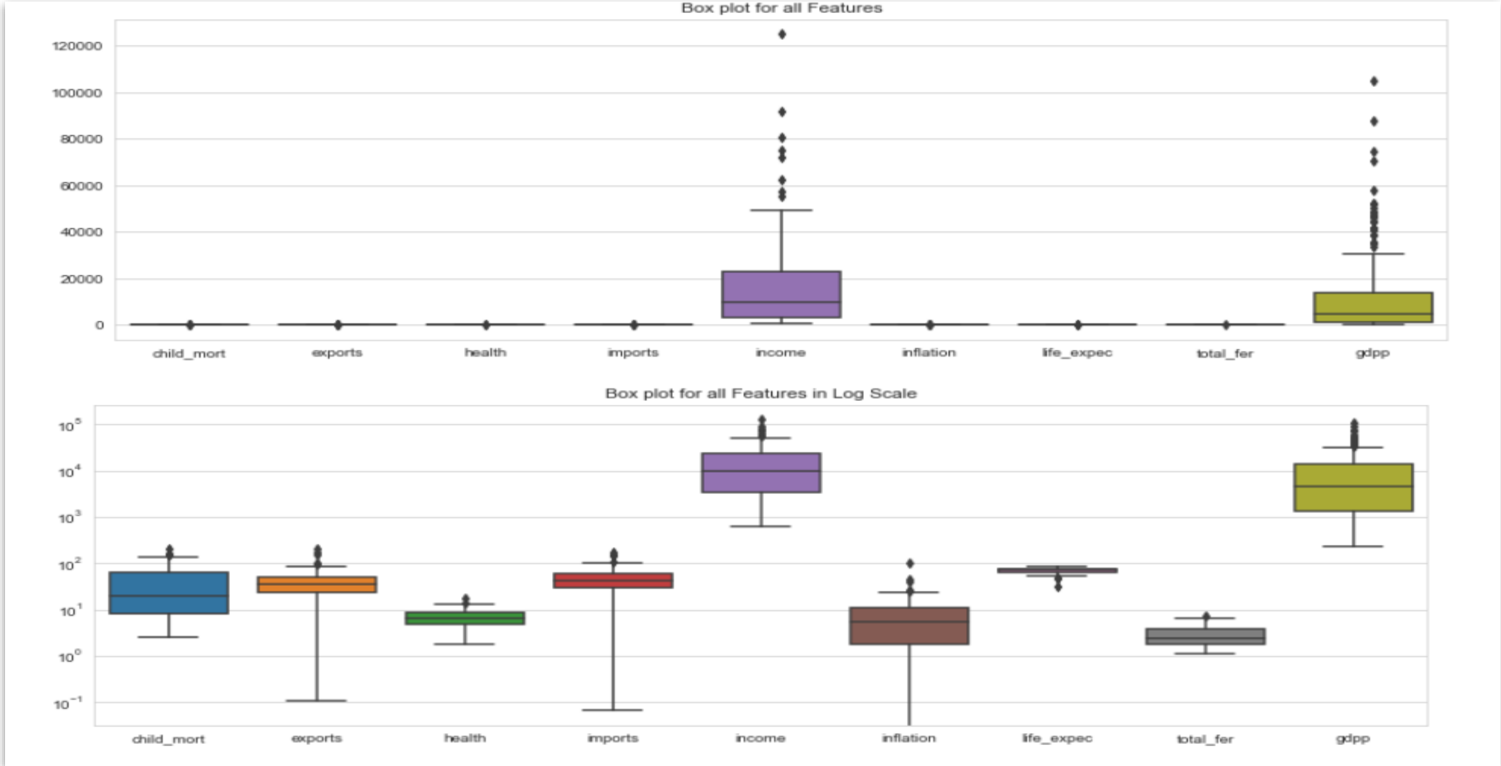
  Q3 + (1.5 * IQR)

- Top 5 Features with most number of outlier

	No of Outlier
gdpp	25
income	8
exports	5
inflation	5
child_mort	4

Exploratory Data Analysis

Outlier Analysis: Below it can be seen the distribution of the Outliers

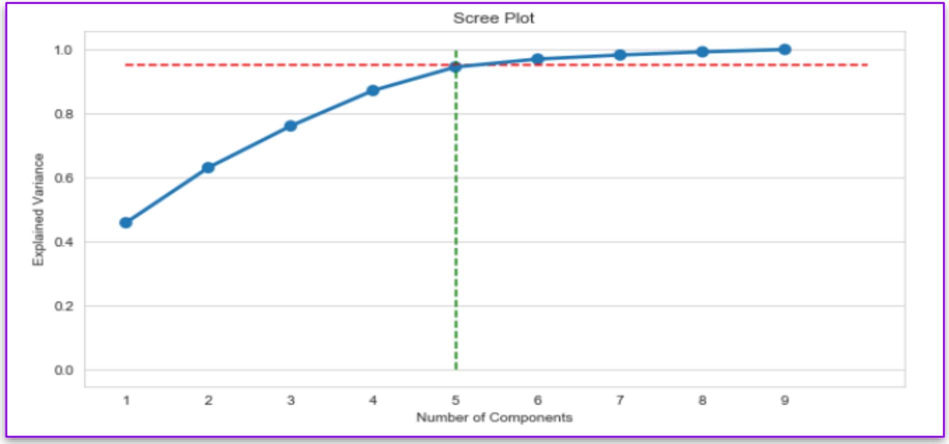


Here the **outliers were not removed**, as then it would result in deleting of **57 Countries** data.

So a decision was taken to look into **outlier** just after **PCA** and **prior to applying Clustering**.

Principal Component Analysis

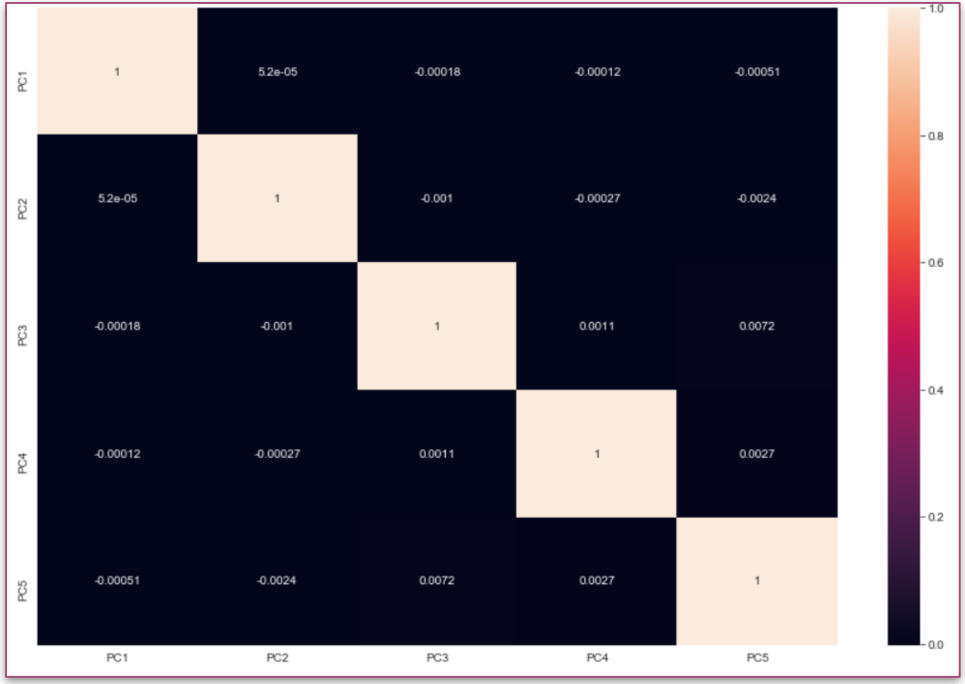
- In the previous slide it was inferred there exists a lot of **multicollinearity**, so now that needs to be removed, Solution is **PCA**
- The approach taken was :
  - First the Feature Scaling was done using the **Standard Scalar**
  - **PCA** was applied on the Scaled Data, **without** specifying the **Number of Components**
  - **Percentage of Variance** explained by each of the **Principal Component** was checked
  - Then using the **Scree plot**, the ideal number of principal components was decided i.e. **5 that captures 95% variance**



Principal Component Analysis

❖ After deciding the number of Principal Component, PCA was applied on the existing Dataset, with the number of component as 5.

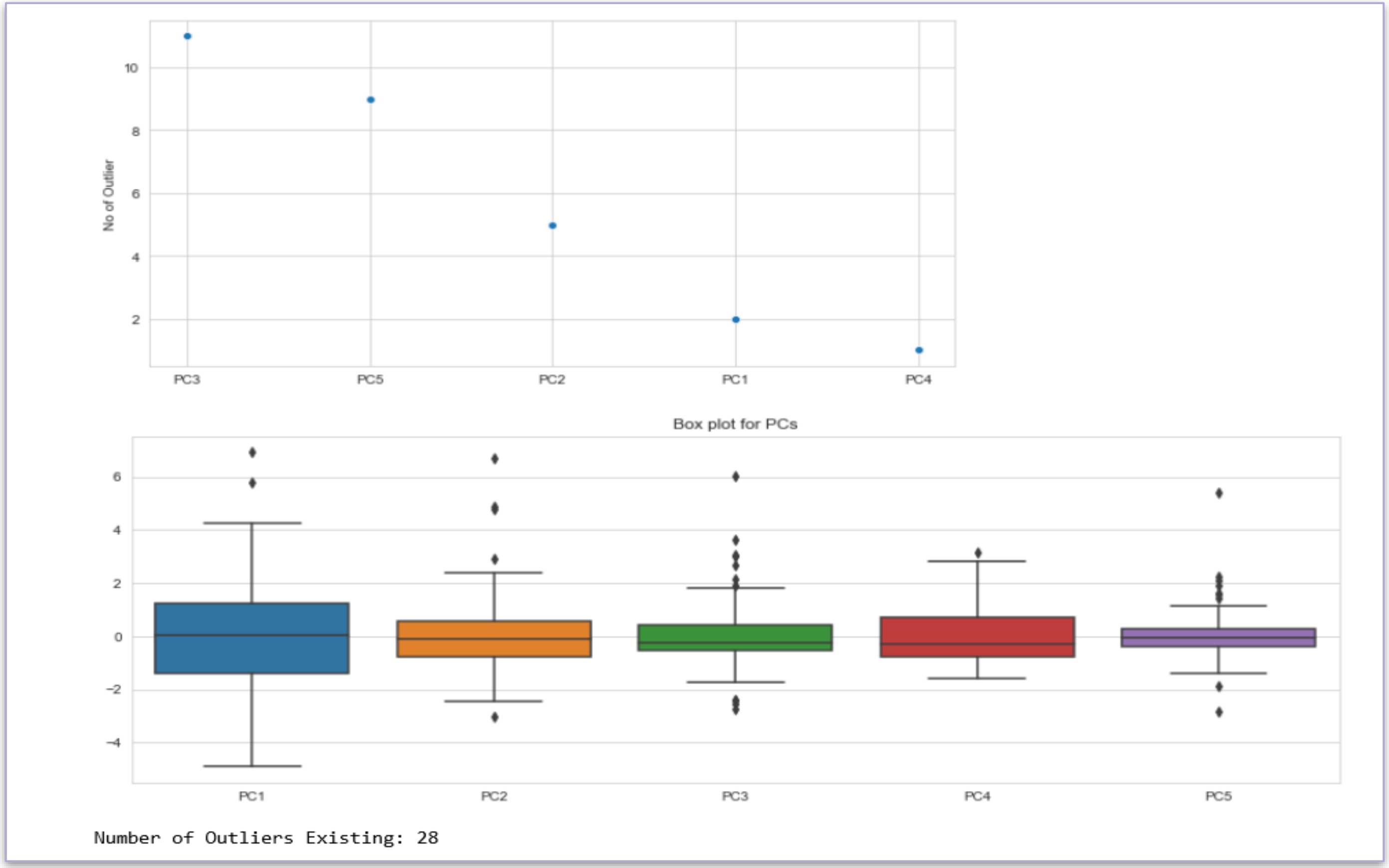
	PC1	PC2	PC3	PC4	PC5
0	-2.913000	0.091969	-0.721242	1.001838	-0.146765
1	0.429870	-0.589373	-0.328611	-1.165014	0.153205
2	-0.285289	-0.452139	1.232051	-0.857767	0.191227
3	-2.932714	1.698771	1.525076	0.855595	-0.214778
4	1.033371	0.133853	-0.216699	-0.846638	-0.193186



- Final shape of the new Dataframe **167** rows , **5** columns
- As it can be seen on the right **no Multicollinearity** exists after applying PCA.

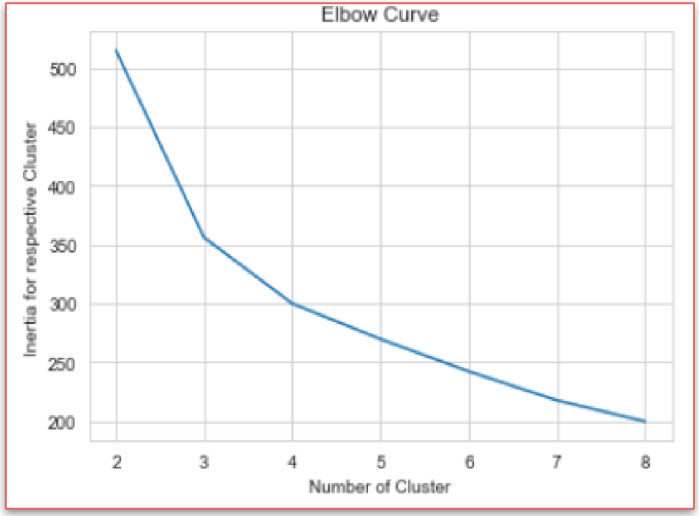
Outlier Removal

- As we can see after applying **PCA,** number of outlier reduced very much hence there will be very less number of Countries data will be removed.
- Hence the Outliers were removed and were identified using **IQR** method.

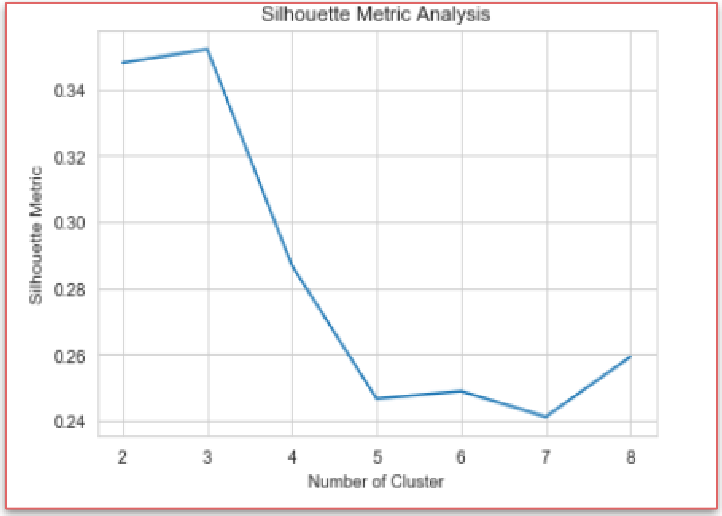


Clustering

- ❖ **Prior** to starting with **Clustering**, **Hopkins Test** is performed to reject the Null Hypothesis which results in **proving the dataset** has a high tendency to cluster.
- ❖ The Result came **> 0.60,** so clustering can be done, as this value is susceptible to change on every run
- ❖ How many cluster Centroid can be chosen ? Soln. **Elbow Curve** and **Silhouette Metric**



Elbow curve, after **cluster 3**, the inertia start decreasing in Linear fashion



At **Cluster 3**, the plot attains the highest values, this implies the data point is very much similar to other data point in cluster

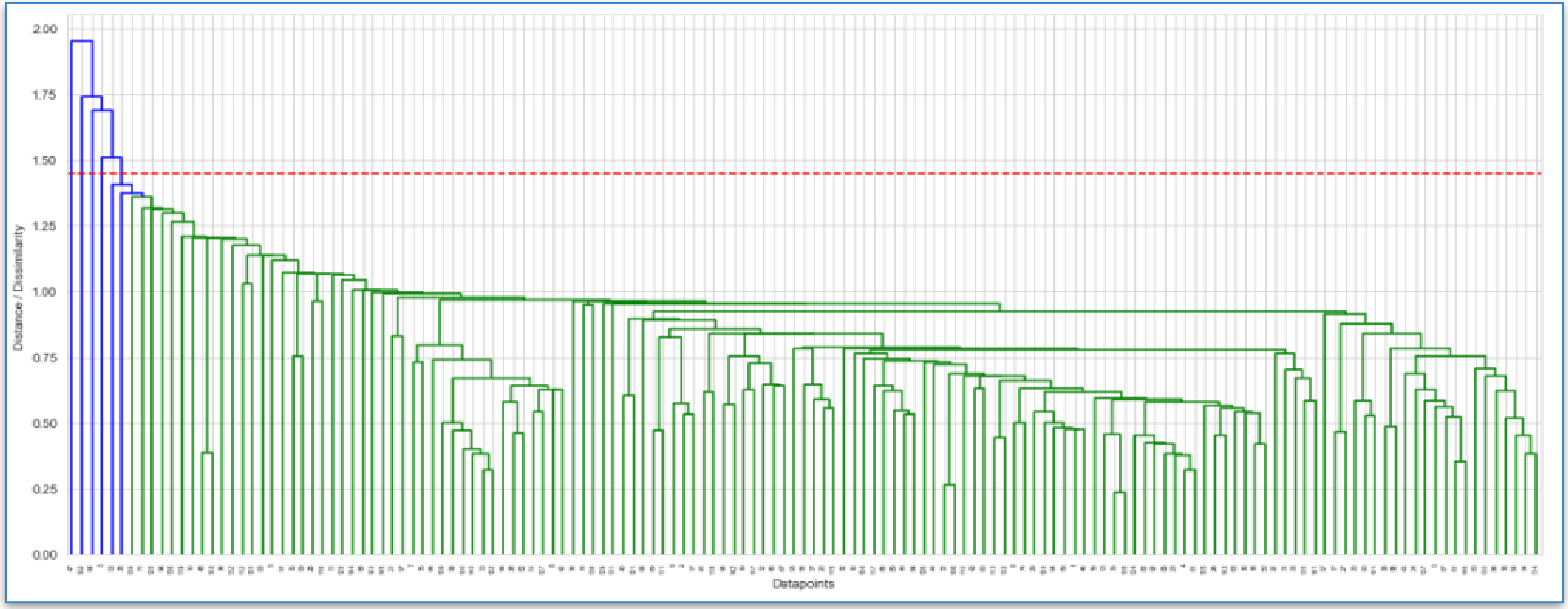
Clustering using K Means

- Using **Number of Cluster** as **3,** Initialization means is used as **K-means ++** so that the centroid are determined using algorithm and not randomly.
- After the applying the KMeans, some of the data in data frame looks like this. Cluster_id represents the Cluster Labels, as determined by K Means.

	country	PC1	PC2	PC3	PC4	PC5	cluster_id
0	Afghanistan	-2.913000	0.091969	-0.721242	1.001838	-0.146765	1
1	Albania	0.429870	-0.589373	-0.328611	-1.165014	0.153205	0
2	Algeria	-0.285289	-0.452139	1.232051	-0.857767	0.191227	0
3	Angola	-2.932714	1.698771	1.525076	0.855595	-0.214778	1
4	Antigua and Barbuda	1.033371	0.133853	-0.216699	-0.846638	-0.193186	0

Clustering using Hierarchical Clustering

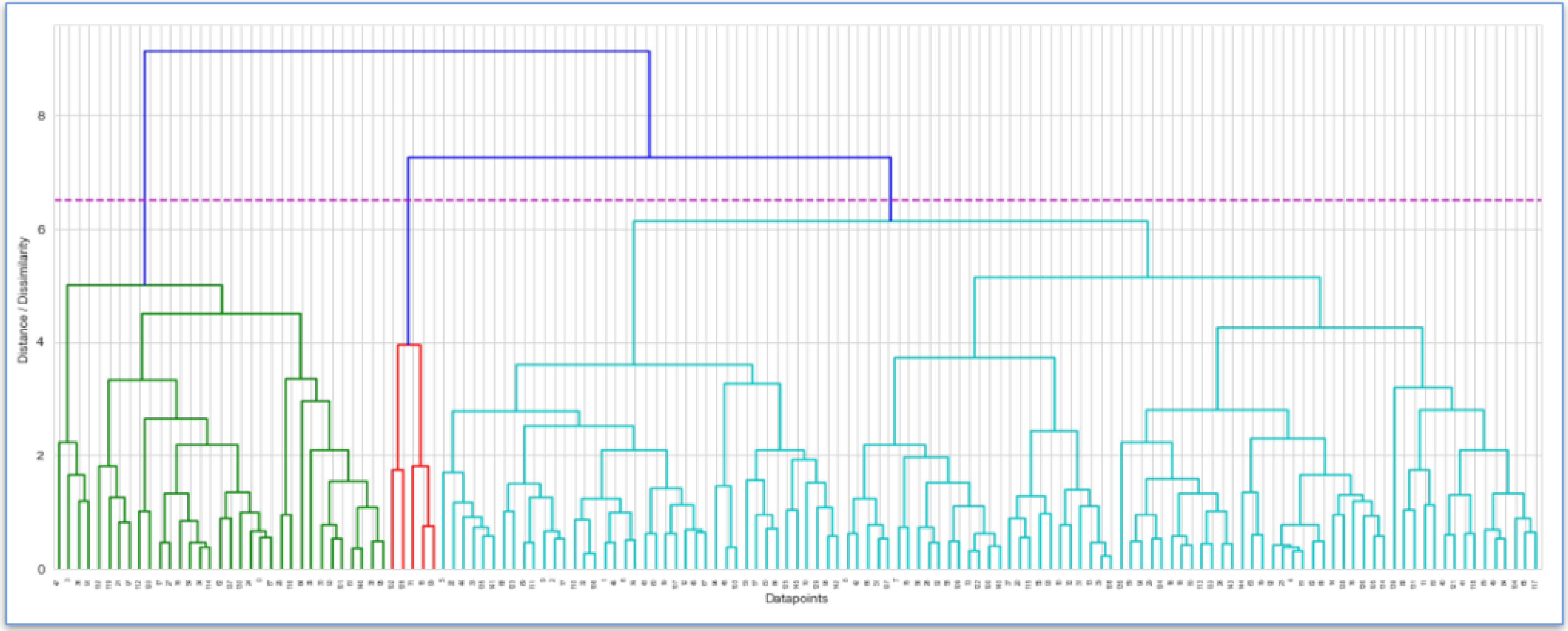
- Using **Single Linkage**, it was considered the Dissimilarity as **Euclidean,** and the **Dendrogram** was constructed.



There are very few clusters are determined. It looks quite crowded in the bottom, lots of horizontal branches on the same height. A decision was taken to take cut off at **1.45**, so that no. of cluster can be **5**

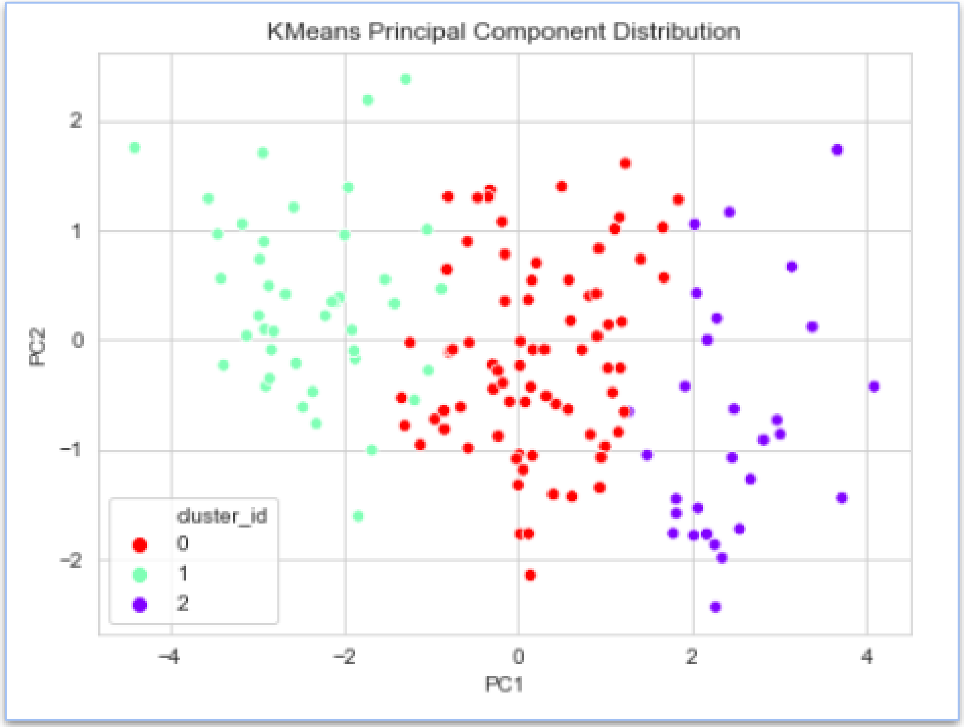
Clustering using Hierarchical Clustering

- Using **Complete Linkage**, it was considered the Dissimilarity as **Euclidean,** and the **Dendrogram** was constructed.



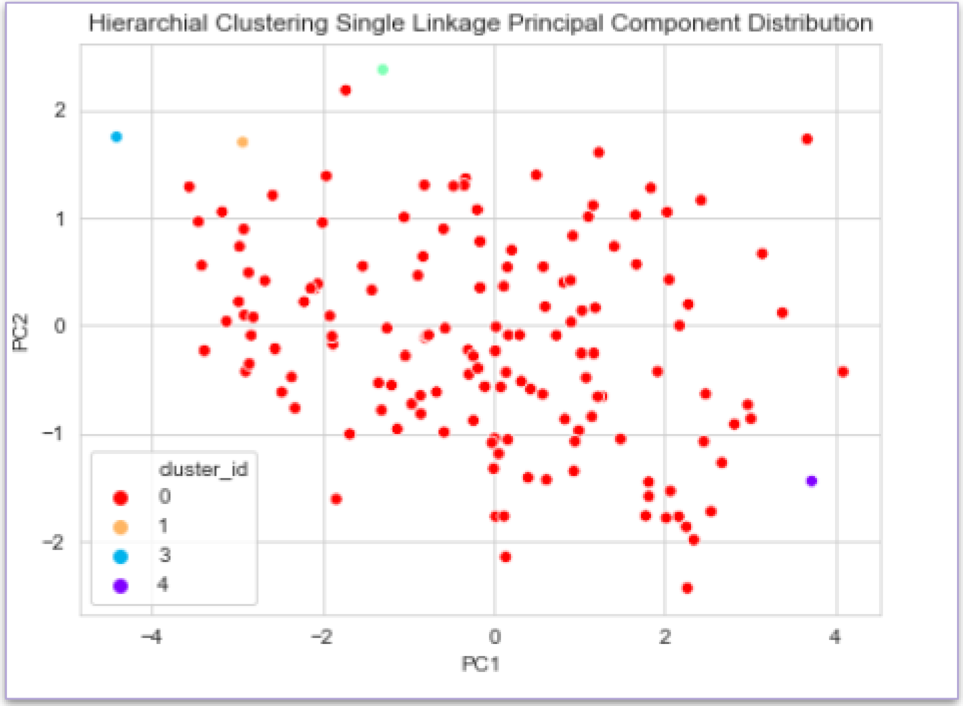
There are many distinct clusters that are determined. A decision was taken to take cut off at **6.5**, so that no. of cluster can be **3**

Visualization of Principal Component w.r.t Cluster IDs



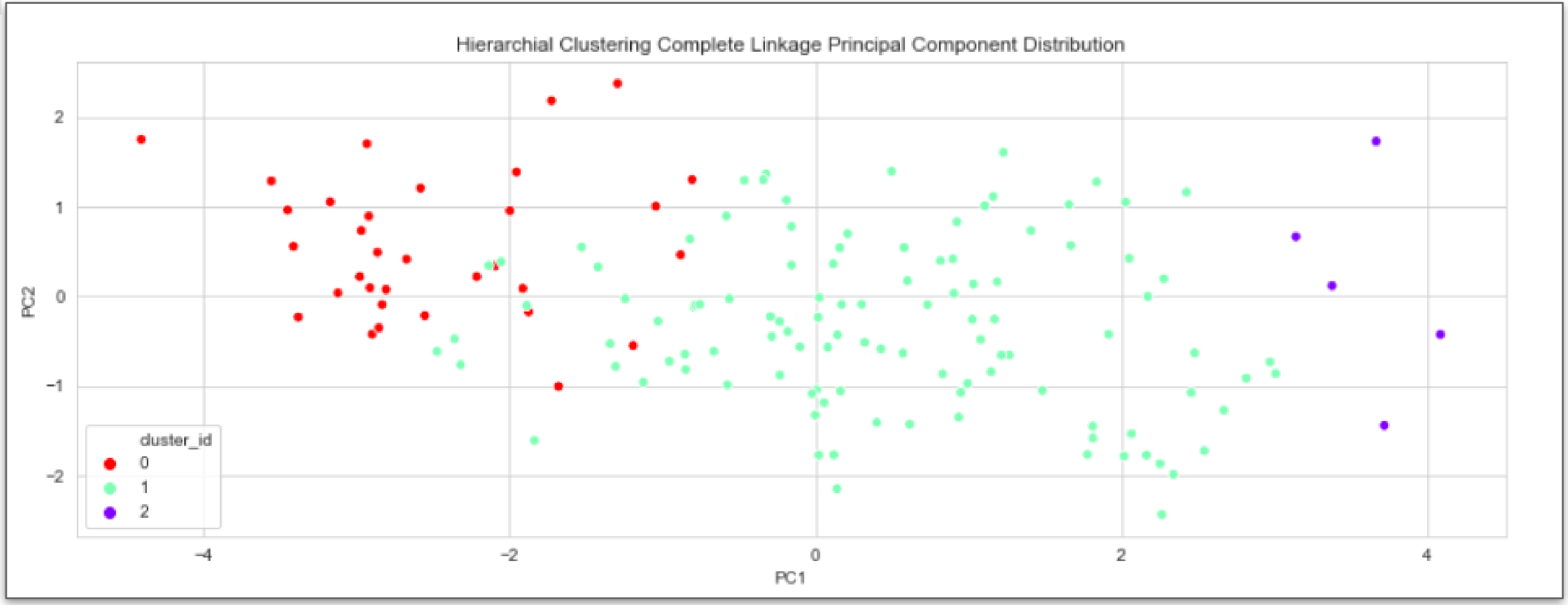
For the **K Means clustering** result we can see that the number of cluster is **3** and they are almost **distinguishable** and **separated**

Visualization of Principal Component w.r.t Cluster IDs



For the **Single Linkage Hierarchical Clustering** most of the data points are clubbed under one cluster, and we can see the cluster **1,2,3,4**, they have identified each cluster as **one data point**

Visualization of Principal Component w.r.t Cluster IDs



For **Complete Linkage Hierarchical Clustering**, cluster 1 has covered the center part , and here also we can see all the clusters are distinguished separately

Clustering Continued...

So a Final Decision was taken to Continue with **K Means** and **Complete Linkage Hierarchical Clustering** for further Analysis

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	cluster_id
0	Afghanistan	90.2	10.0	7.58	44.9	1610	9.44	56.2	5.82	553	1
1	Albania	16.6	28.0	6.55	48.6	9930	4.49	76.3	1.65	4090	0
2	Algeria	27.3	38.4	4.17	31.4	12900	16.10	76.5	2.89	4460	0
3	Angola	119.0	62.3	2.85	42.9	5900	22.40	60.1	6.16	3530	1
4	Antigua and Barbuda	10.3	45.5	6.03	58.9	19100	1.44	76.8	2.13	12200	0

KMeans Final Dataframe

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	cluster_id
0	Afghanistan	90.2	10.0	7.58	44.9	1610	9.44	56.2	5.82	553	0
1	Albania	16.6	28.0	6.55	48.6	9930	4.49	76.3	1.65	4090	1
2	Algeria	27.3	38.4	4.17	31.4	12900	16.10	76.5	2.89	4460	1
3	Angola	119.0	62.3	2.85	42.9	5900	22.40	60.1	6.16	3530	0
4	Antigua and Barbuda	10.3	45.5	6.03	58.9	19100	1.44	76.8	2.13	12200	1

Complete Linkage Hierarchical Clustering Final Dataframe

Analysis on Clustered Dataframe

Aggregation on **K Means Clustered Data** based Cluster ID

cluster_id	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
0	21.45	40.37	6.17	47.09	11915.66	6.72	73.03	2.29	6138.61
1	91.44	29.98	5.95	39.14	4197.98	10.20	59.56	4.96	1951.83
2	4.66	46.32	9.46	44.09	38058.62	1.55	80.16	1.75	40410.34

- For cluster **2**: It's more like a **Developed Country**, where child_mort,inflation,total_fert is less, exports, healthimports, income, life_expec and gdpp is more
- For cluster **0**: it's more like a **Developing country** where child_mort, income, life_expec and gdpp is less compared to Developed countries,
- For cluster **1**: They are more like a **Underdeveloped countries** that needs attention, child_mort is quite high, income is also very low, gdpp & life_expec is also very low

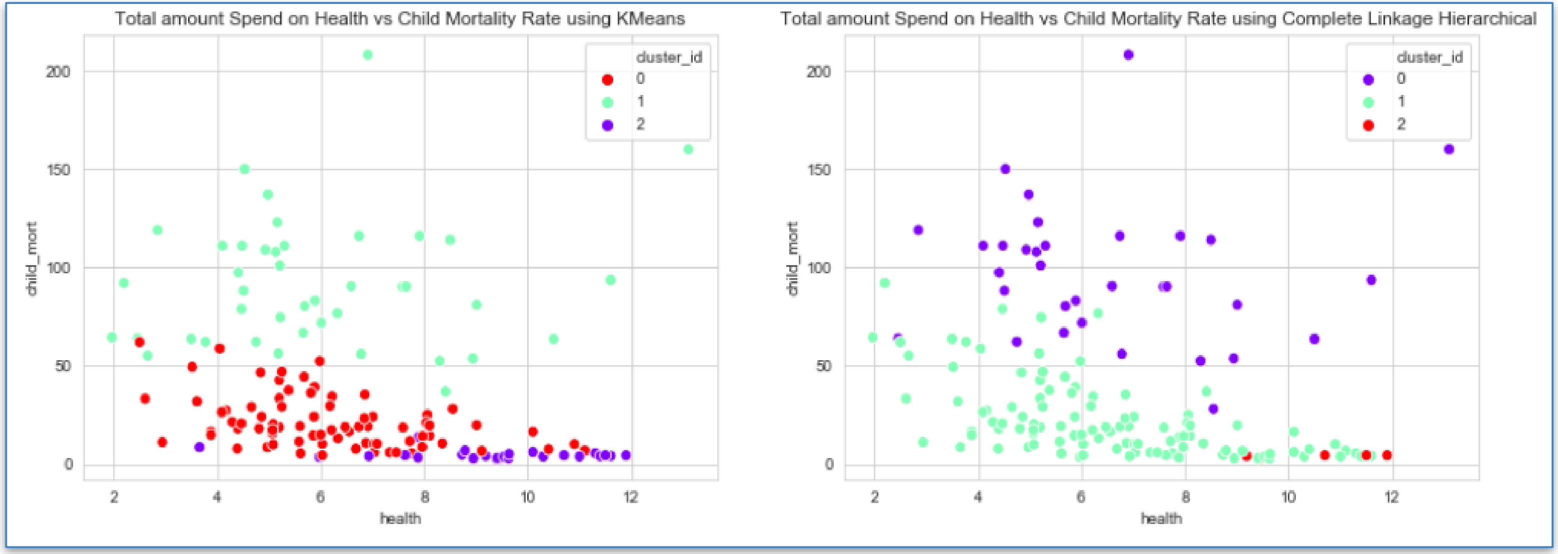
Analysis on Clustered Dataframe

Aggregation on **Hierarchical Clustered Data** based Cluster ID

cluster_id	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
0	97.19	31.92	6.51	43.63	3766.52	8.93	58.14	5.20	1874.36
1	21.81	39.10	6.66	43.63	16616.61	6.27	73.89	2.31	12410.61
2	4.18	71.02	10.55	61.32	50020.00	1.16	80.86	1.83	61160.00

- For cluster **2**, it looks more like a **Developed country**, child_mort is very less, income, exports, health, imports, gdpp is huge
- For cluster **1**, it looks more like a **Developing country**, child mort is less than developed country cluster,
- For Cluster **0**, child_mort is highest of all, exports, health, income, life_expec gdpp is lowest of all cluster, so this seems to be representing the **Underdeveloped country**

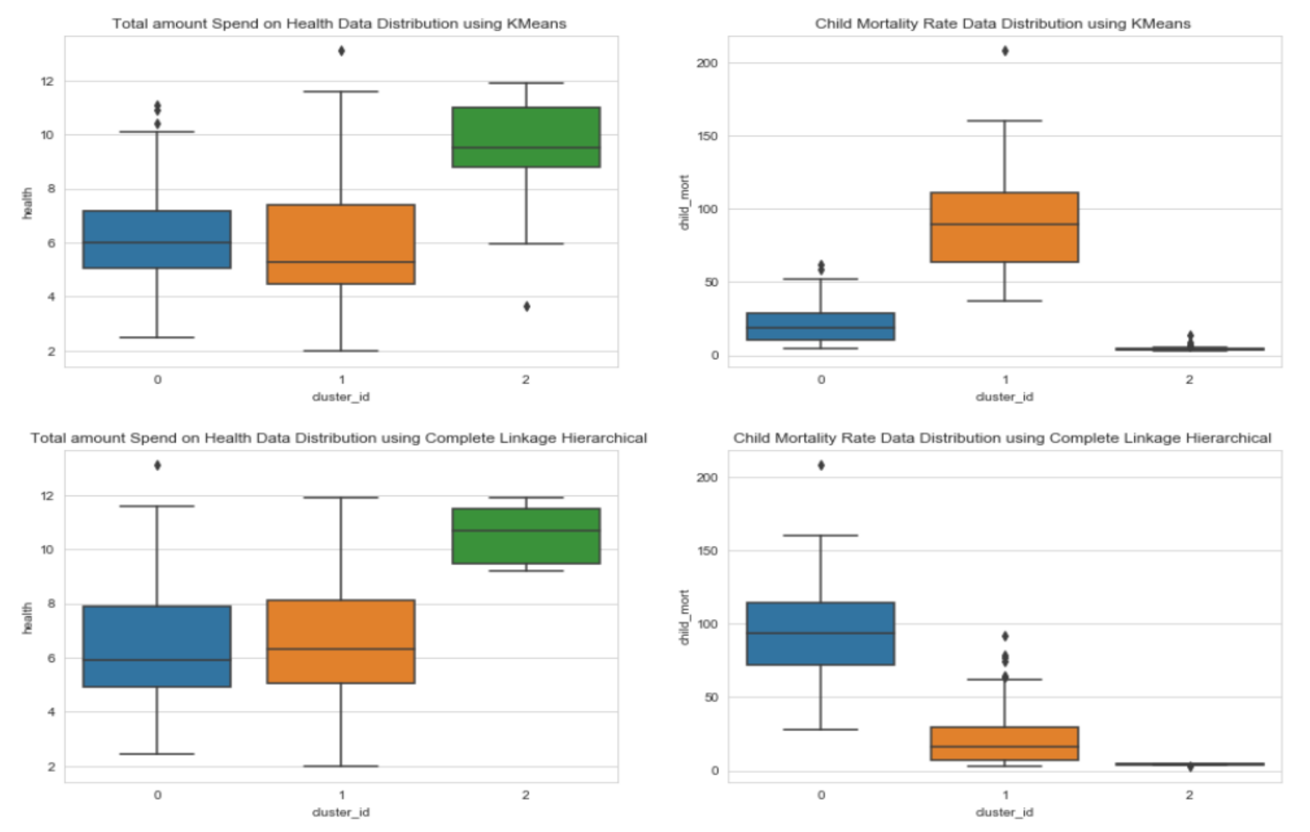
Analysis of child_mort vs health



For **K Means** cluster plot, we can see that Cluster 1 rightly identifies the value which are having **low Health Spending** and **High Child Mortality**

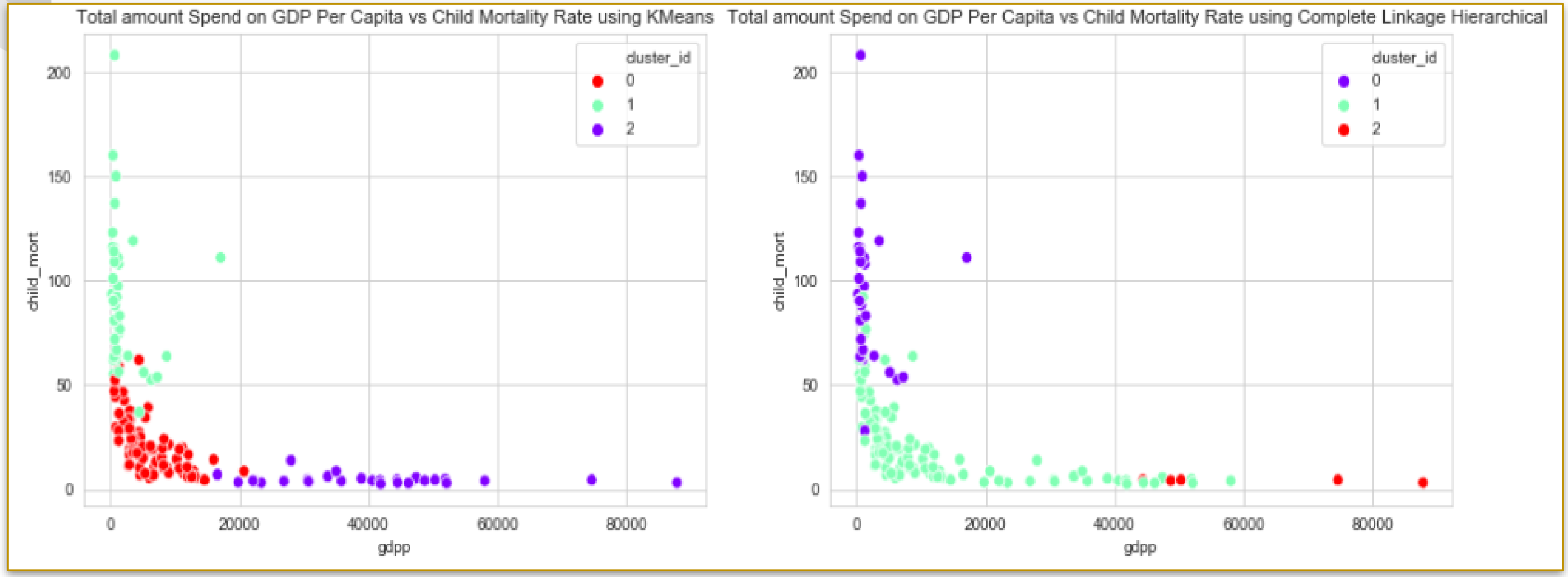
For **Hierarchical Clustering**, **cluster 0** identifies the **low Health Spending and high child mortality**

Analysis of child_mort vs health



- **K Means**: For Health, **Cluster 1** is capturing all the countries which has lower spending towards health as the lower whisker and also the median is lowest and for Child Mortality rate, cluster 1 is capturing the highest child mortality rate
- **Hierarchical Clustering**: Cluster 0 and 1 is capturing the low spending on Health, but for Child Mortality rate, Cluster 0 is capturing the high spending on Child mortality rate

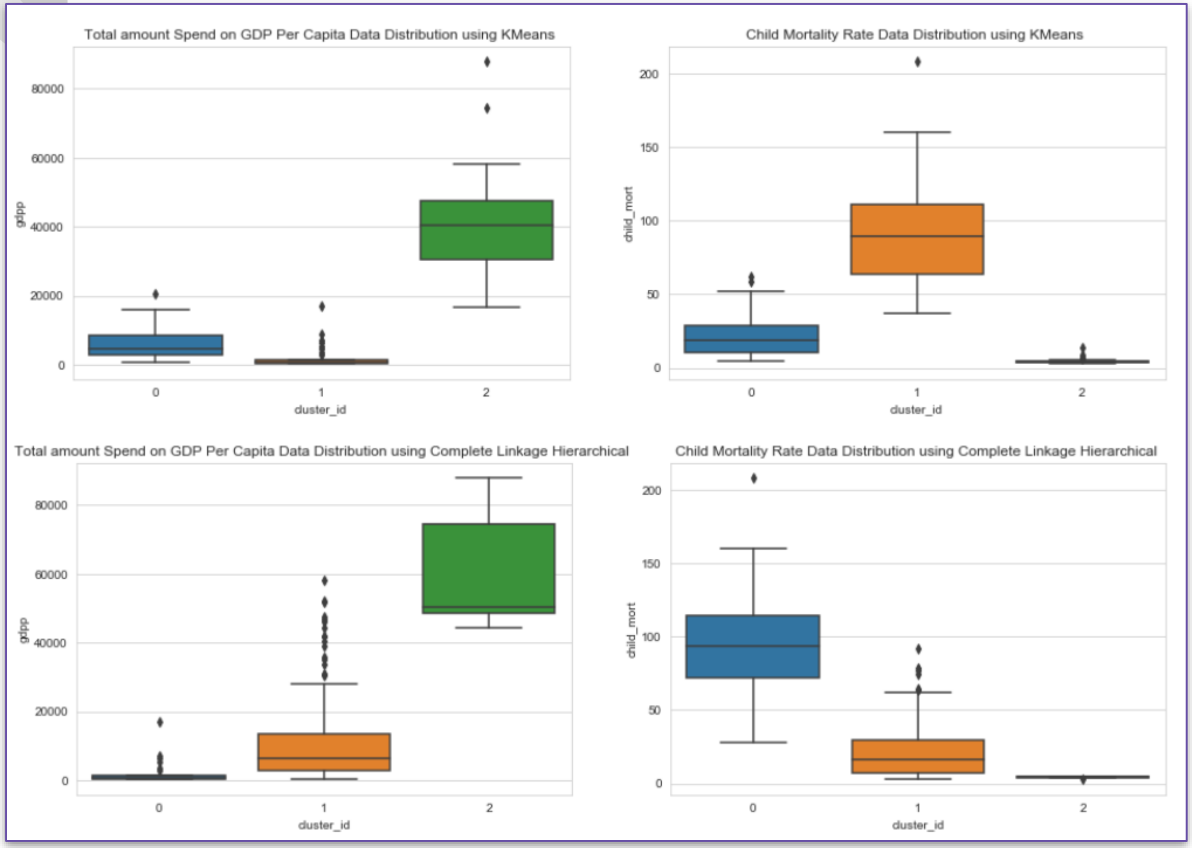
Analysis of gdpp vs child_mort



For the **K Means Clustering** , cluster 1 correctly identifies the where there is **low GDP** and **High Child Child Mortality**

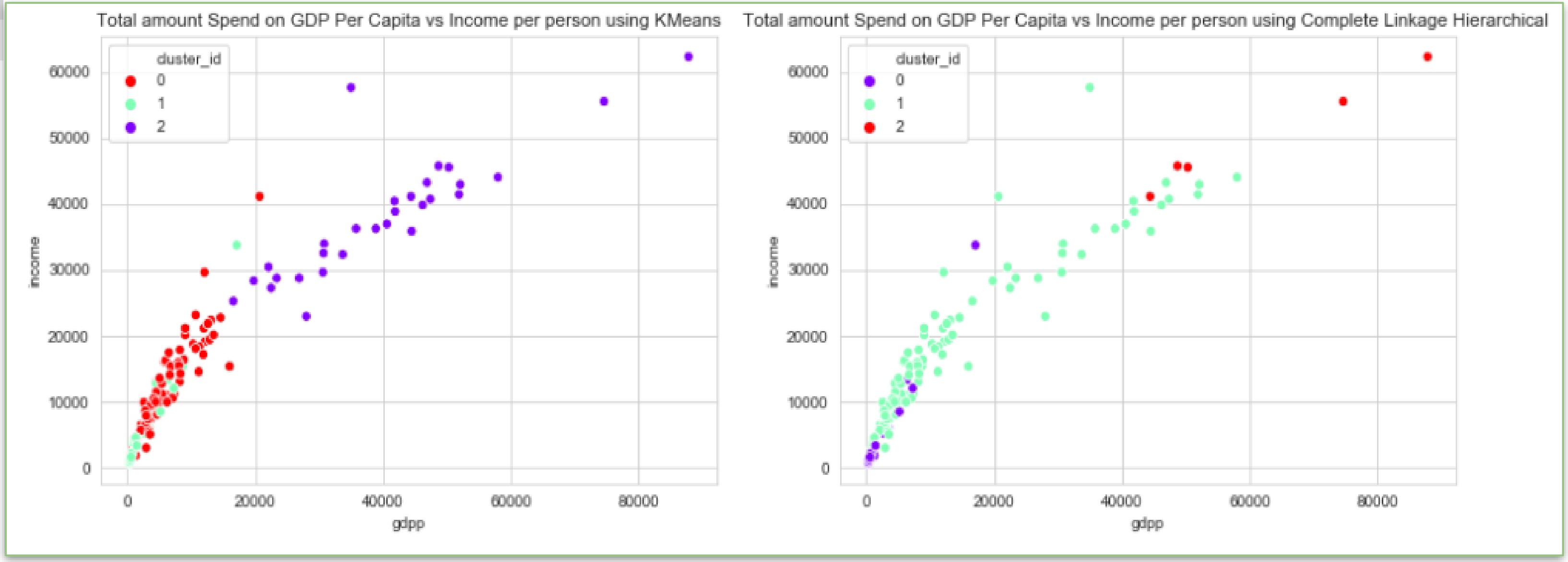
For the **Complete Linkage Hierarchical** Clustering , cluster 0 correctly identifies the where there is **low GDP and High Child Child Mortality**

Analysis of gdpp vs child_mort



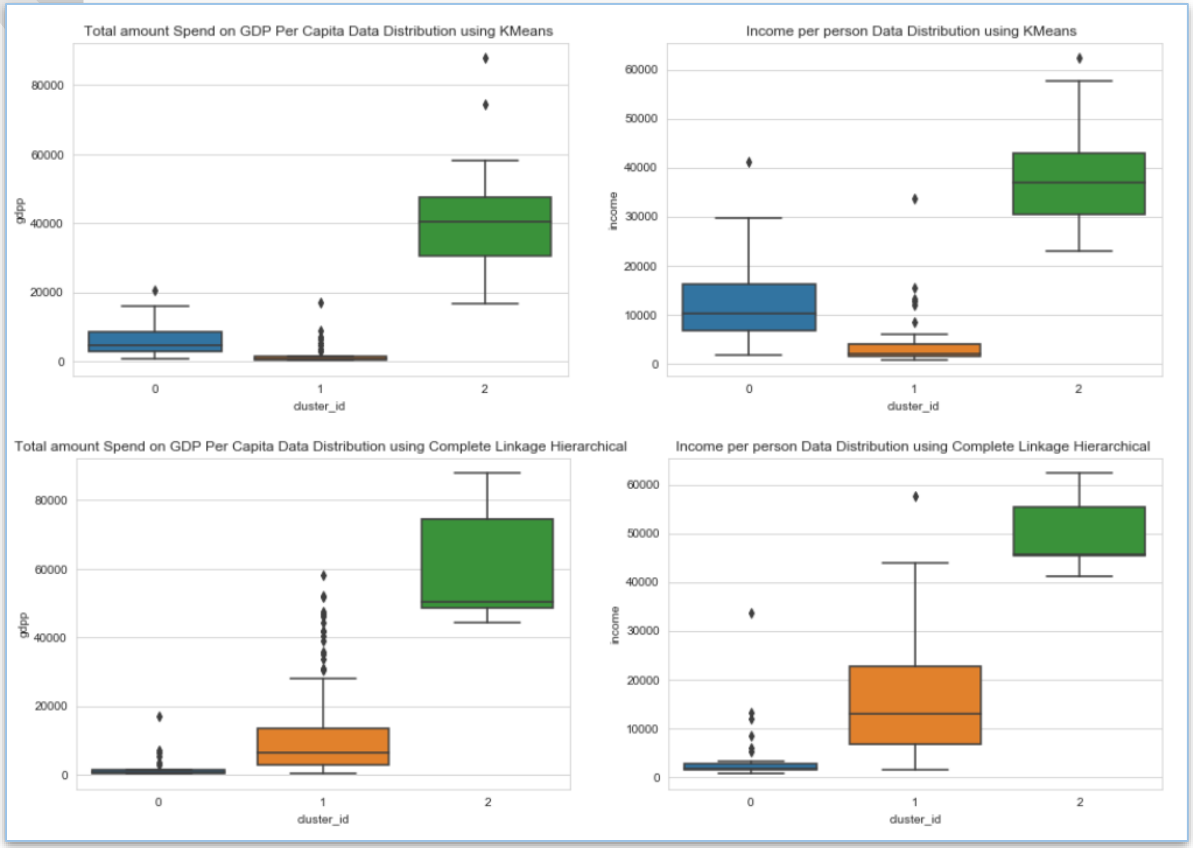
- **K Means**: For gdpp, cluster 1 capturing all the lower values, and as above plot, for spending on health Cluster 1 is capturing the highest spending on health
- **Hierarchical Clustering**: Cluster 0 is capturing the low spending on gdpp, but for child mortality cluster 0 is capturing the highest value

Analysis of gdpp vs income



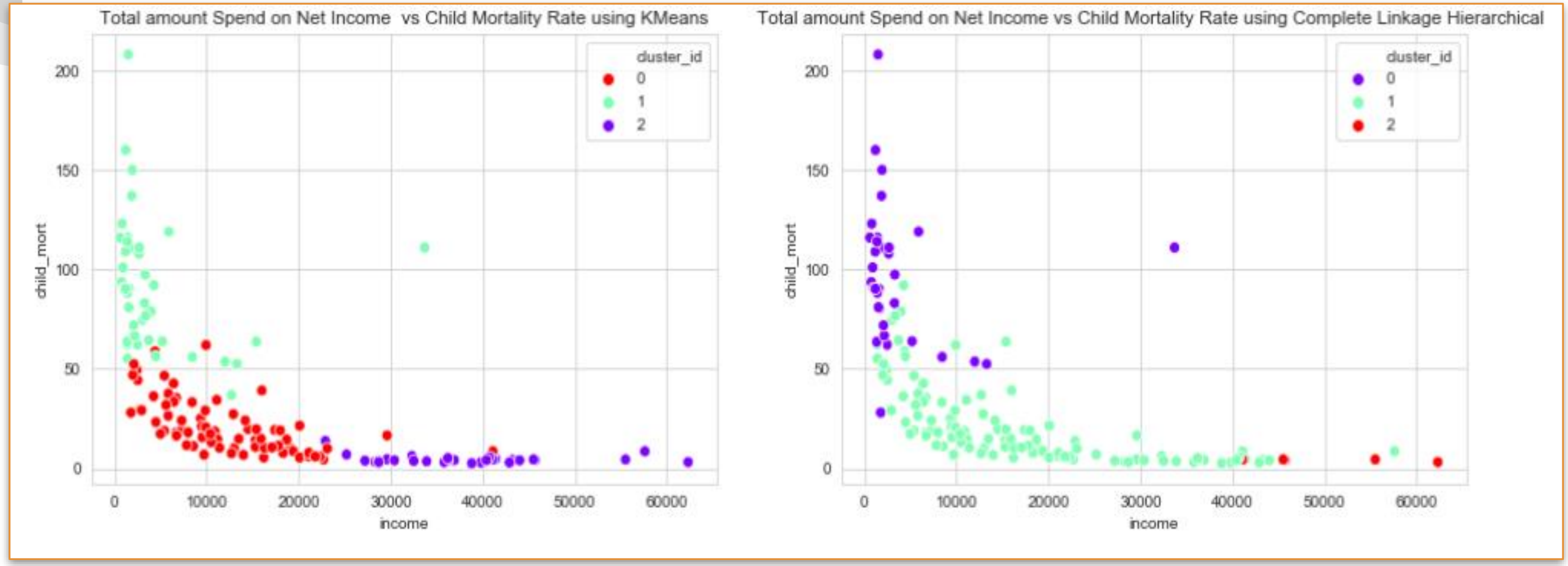
For **lower income** than 25th percentile and low gdpp than 25th percentile. Cluster 1 identifies properly for **K Means** and for **Hierarchical Clustering** the Cluster 0 identifies properly

Analysis of gdpp vs income



- **K Means**: For gdpp, cluster 1 capturing all the lower values, and cluster 1 is capturing the information specific for the persons having lowest income
- **Hierarchical Clustering**: For gdpp cluster 0 is capturing the lowest spending on gdpp and Cluster 0 is also capturing the lowest income

Analysis of income vs child_mort



As we can see for the data **less than 25th percentile of Net Income** and **more than 75th Percentile Child Mortality Rate** is identified by Cluster 1 in **K Means** and Cluster 0 for **Hierarchical Cluster using Complete Linkage**.

Decision Taken

- So the Final Conclusion is From **K Means Cluster 1** provided information, that will help us in getting the country list who needs most number of attention and from **Hierarchical Clustering** the **Cluster 0** is capturing the **lowest information**. So now we would try to derive the list of countries who needs most amount of attention

- And as we can see as most of the results for **K Means** is **almost same** like **Hierarchical clustering**. So taking a decision here to go with **KMeans**

- Features that were being selected using which a conclusion can be driven.
    - **Health** : Capturing the **lowest** spend
    - **Child Mortality Rate** : Capturing the **highest** value of child mortality
    - **GDP Per Capita** : Capturing the **lowest** GDP per capita
    - **Net Income per person** : Capturing the **lowest** income

- Now for all of the **Health** and **Child mortality** we would get all the countries belonging **Cluster 1** and whose values are **less** than **mean** for **health** and **more** than **mean** for **child mortality rate** (why because they have very less outliers)

- For GDP Per capita and Income we would get all the countries belonging **Cluster 1** and whose values are **less** than **10th Quantile**

Conclusion

So from the previous taken decision, we got the top 10 Countries which requires most amount of attention based on **socio-economic and health factors**

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
24	Burkina Faso	116.0	19.20	6.74	29.6	1430	6.81	57.9	5.87	575
25	Burundi	93.6	8.92	11.60	39.2	764	12.30	57.7	6.26	231
35	Congo, Dem. Rep.	116.0	41.10	7.91	49.6	609	20.80	57.5	6.54	334
48	Eritrea	55.2	4.79	2.66	23.3	1420	11.60	61.7	4.61	482
61	Guinea	109.0	30.30	4.93	43.2	1190	16.10	58.0	5.34	648
62	Guinea-Bissau	114.0	14.90	8.50	35.2	1390	2.97	55.6	5.05	547
86	Madagascar	62.2	25.00	3.77	43.0	1390	8.79	60.8	4.60	413
95	Mozambique	101.0	31.50	5.21	46.2	918	7.64	54.5	5.56	419
101	Niger	123.0	22.20	5.16	49.1	814	2.55	58.8	7.49	348
116	Sierra Leone	160.0	16.80	13.10	34.5	1220	17.20	55.0	5.20	399

'Burkina Faso', 'Burundi', 'Congo, Dem. Rep.', 'Eritrea', 'Guinea', 'Guinea-Bissau', 'Madagascar', 'Mozambique', 'Niger', 'Sierra Leone'