Submitted for Upgrad Data Science Program - 2019

ASHOK GORANTALA

CLUSTERING & PCA ASSIGNMENT

Dated: 11th Nov 2019

PROBLEM STATEMENT:

- Perform PCA on the dataset and obtain the new dataset with the Principal Components. Choose the appropriate number of components k. You need to perform your clustering activity on this new dataset, i.e. the PCA modified dataset with the k components.
- **Outlier Analysis:** You must perform the Outlier Analysis on the dataset, before or after performing PCA, as per your choice. However, you do have the flexibility of not removing the outliers if it suits the business needs or a lot of countries are getting removed. Hence, all you need to do is find the outliers in the dataset, and then choose whether to keep them or remove them depending on the results you get.
- Try both **K-means** and **Hierarchical** clustering(both single and complete linkage) on this dataset to create the clusters. [Note that both the methods may not produce identical results and you might have to choose one of them for the final list of countries.]
- Analyse the clusters and identify the ones which are in dire need of aid. You can analyse the clusters by comparing how these three variables - [**gdpp, child_mort and income**] vary for each cluster of countries to recognise and differentiate the clusters of developed countries from the clusters of under-developed countries. Note that you perform clustering on the PCA modified dataset and the clusters that are formed are being analysed now using the original variables to identify the countries which you finally want to select.
- Also, you need to perform visualisations on the clusters that have been formed. You can do this by choosing the first two Principal Components (on the X-Y axes) and plotting a scatter plot of all the countries and differentiating the clusters. You should also do the same visualisation using any two of the original variables (like gdpp, child_mort, etc.) on the X-Y axes as well. You can also choose other types of plots like boxplots, etc.
- The final list of countries depends on the number of components that you choose and the number of clusters that you finally form. Also, both K-means and Hierarchical may give different results. Hence, there might be some subjectivity in the final number of countries that you think should be reported back to the CEO. Here, make sure that you report back at least 5 countries which are in direst need of aid from the analysis work that you perform.

DATA OUTLIER VERIFICATION

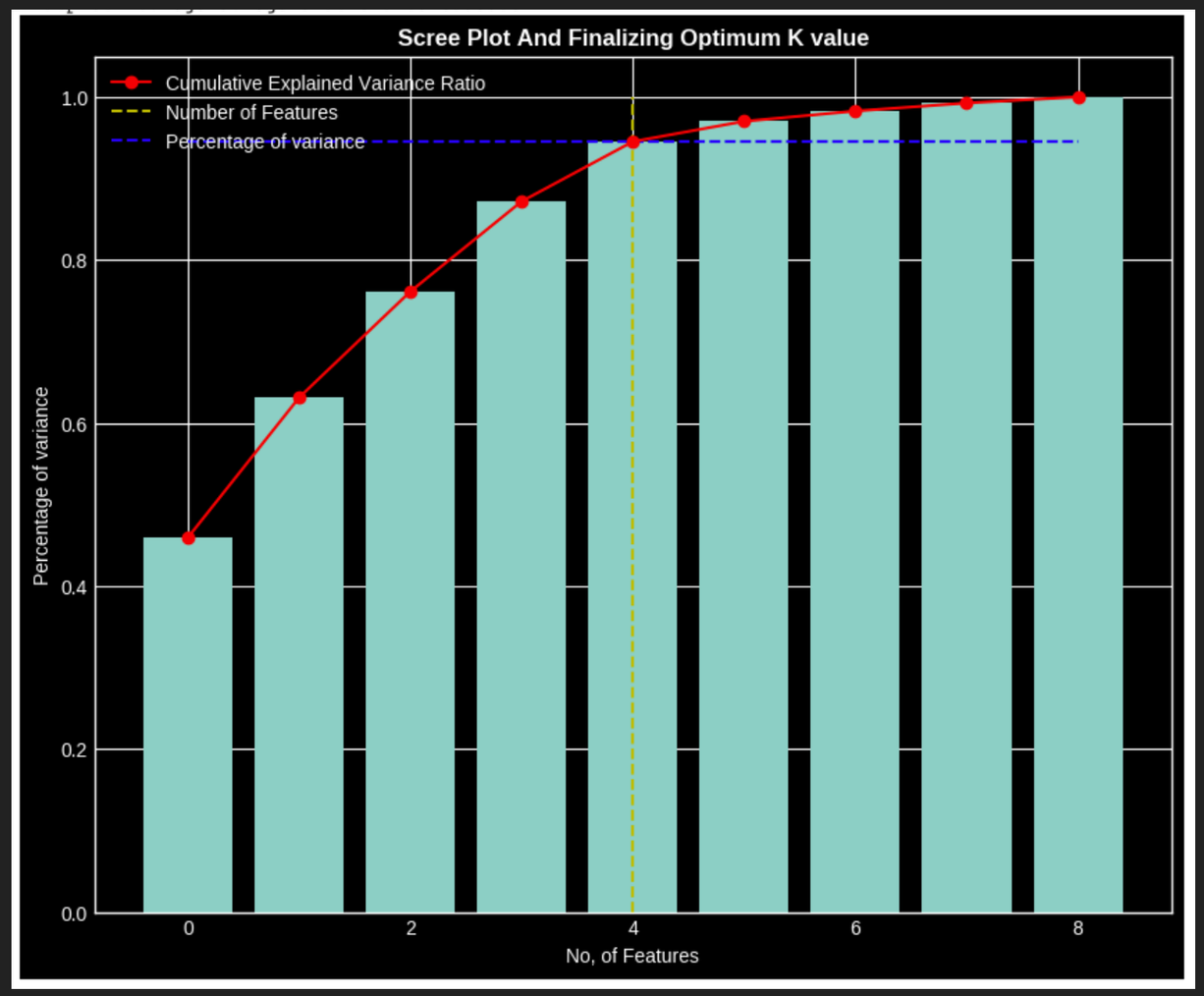
	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
count	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000
mean	38.270060	41.108976	6.815689	46.890215	17144.688623	7.781832	70.555689	2.947964	12964.155689
std	40.328931	27.412010	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
85%	88.310000	65.330000	9.631000	64.520000	36200.000000	15.330000	79.910000	5.023000	31090.000000
90%	100.220000	70.800000	10.940000	75.420000	41220.000000	16.640000	80.400000	5.322000	41840.000000
95%	116.000000	80.570000	11.570000	81.140000	48290.000000	20.870000	81.400000	5.861000	48610.000000
97%	130.140000	87.038000	11.802000	92.768000	62496.000000	23.626000	81.902000	6.230600	52218.000000
98%	145.160000	100.056000	11.900000	105.760000	74208.000000	25.988000	82.000000	6.450400	66364.000000
99%	153.400000	160.480000	13.474000	146.080000	84374.000000	41.478000	82.370000	6.563600	79088.000000
max	208.000000	200.000000	17.900000	174.000000	125000.000000	104.000000	82.800000	7.490000	105000.000000

From the above data observations:

* ’inflation', 'gdpp', 'income' has a sudden spike at 99% and most of the features are gradually increasing too.

PRINCIPAL COMPONENT ANALYSIS

- Upon working with recalling of the data, whilst excluding Country column, we have performed StandardScaling over the numerical data columns.
- Upon getting the “Re-scalled” data, we have used PCA method to check the ‘variance ratio’ for each column and tried finding the columns explaining maximum variance of the data. This is explained by ‘Scree’ plot as below

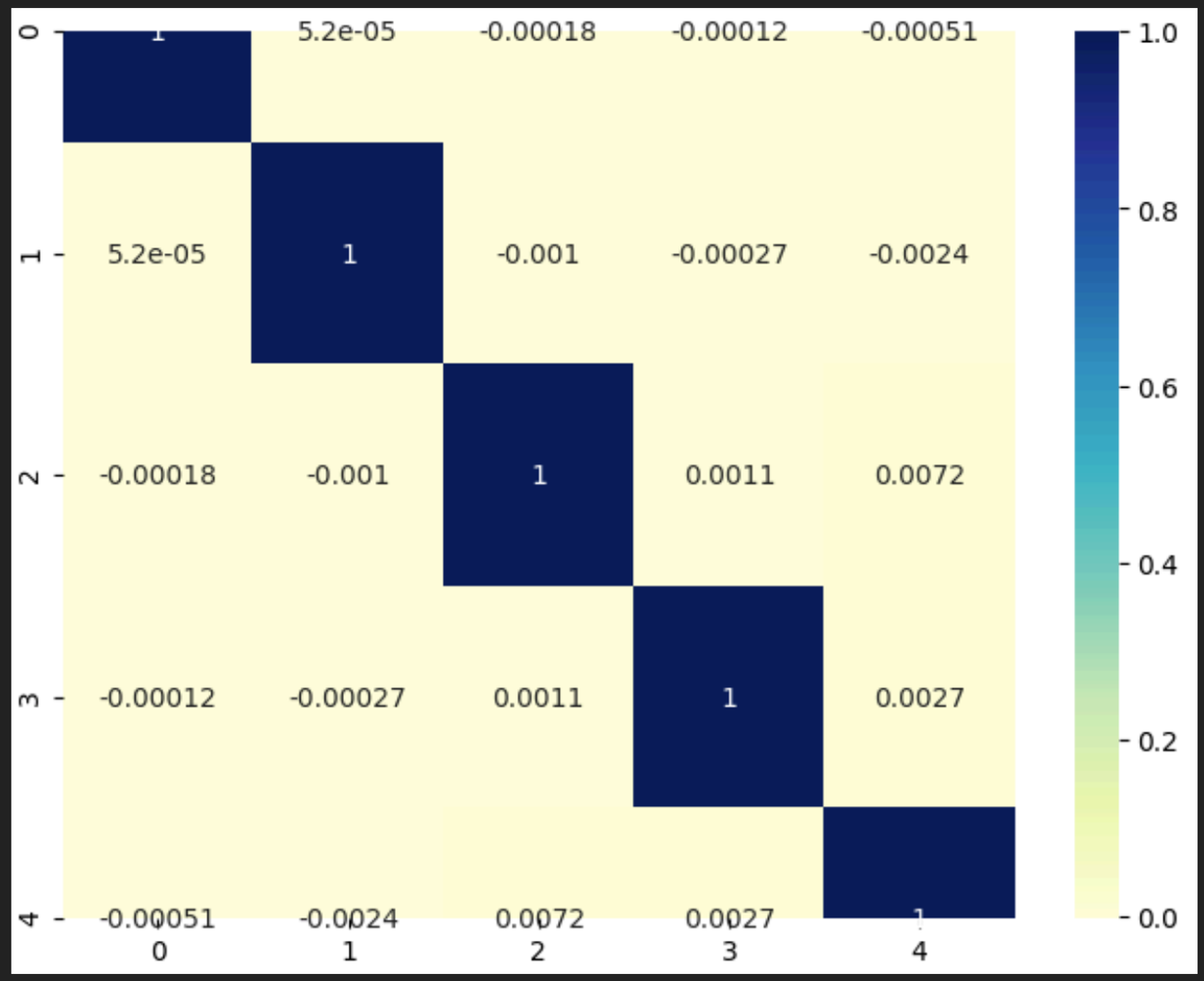


Observations:

1. As per above plot, 5 Features are enough to explain the 94.5 percentage of variance in data
2. selecting 5 features to assess the clustering methodologies.

PRINCIPAL COMPONENT ANALYSIS…CONT

- With the confirmation of having 5 Features, we selected and formed Eighen Values and Eighen Vectors. Then figured the inter-feature dependency aka,. Multicolinearity explained by correlation coeffienct matrix by following plot.



- As expected that PCA helps in reducing the multi-collinearity, we have 5 features strictly having No collinearity and having maximum variance of the total data explained by plain 5 features.

- Where as the data frame head post the creation of NEW dimensioned data, looks like:

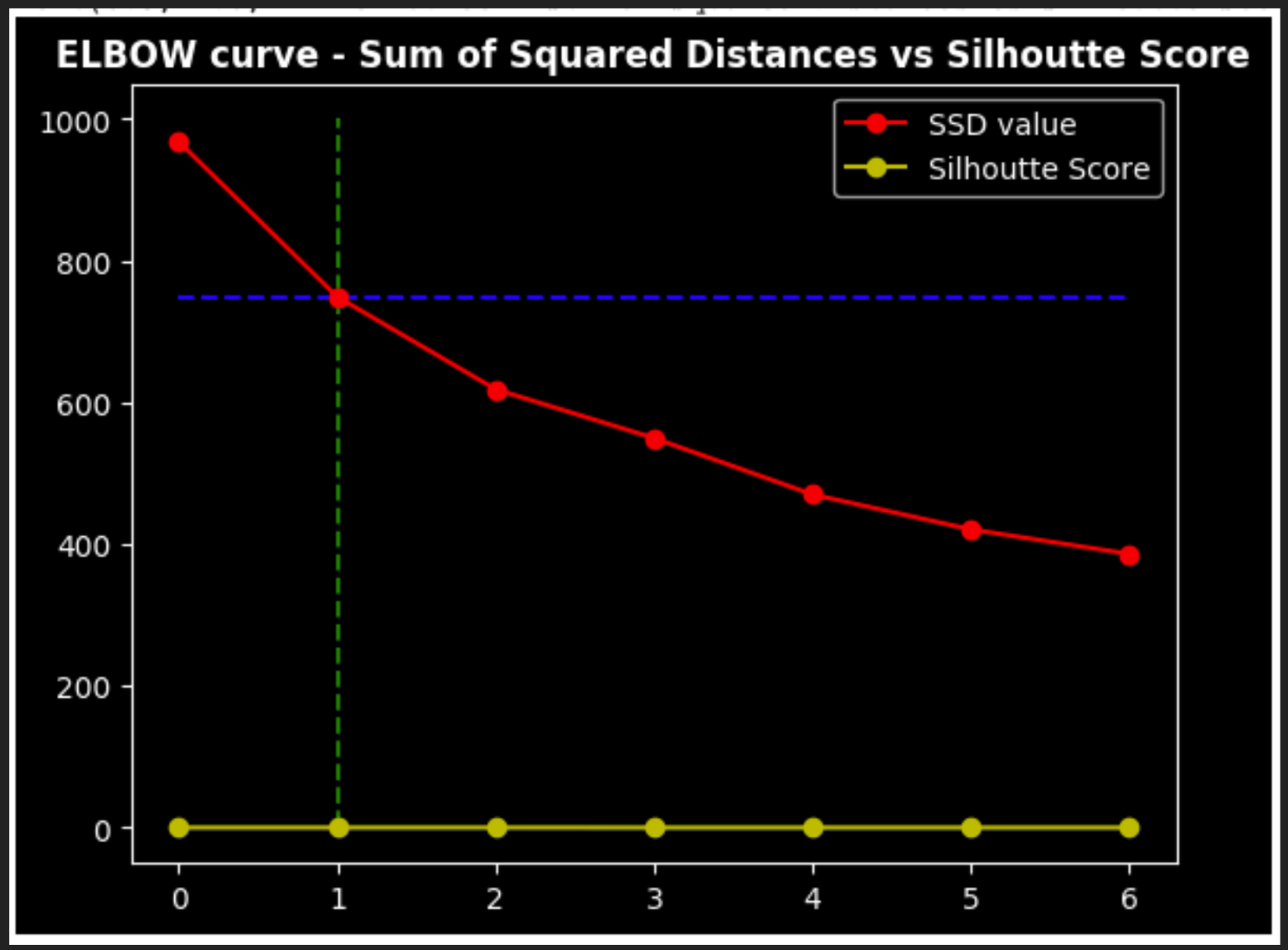
	PC1	PC2	PC3	PC4	PC5
0	-2.913000	0.091969	-0.721242	1.001838	-0.146765
1	0.429870	-0.589373	-0.328611	-1.165014	0.153205
2	-0.285289	-0.452139	1.232051	-0.857767	0.191227
3	-2.932714	1.698771	1.525076	0.855595	-0.214778
4	1.033371	0.133853	-0.216699	-0.846638	-0.193186

PRINCIPAL COMPONENT ANALYSIS…CONT

- **Hopkins Rating:**
  - This parameter signifies the strength of inter-cluster distance compared with intra-cluster distance. That being, having high-Hopkins rating is considered to be having high probability of creating better clusters.
  - And, for the PCAed data set, we have Hopkins rating ranging from **0.77** to **0.87**. This is again considered to be very good value to have high probability of creating clusters.

CLUSTERING – K MEANS METHODOLOGY

- **Finding Optimal number of Clusters - K Value:**
  - **Using the SSD (Sum of Squared Distances) and Silhoutte score**

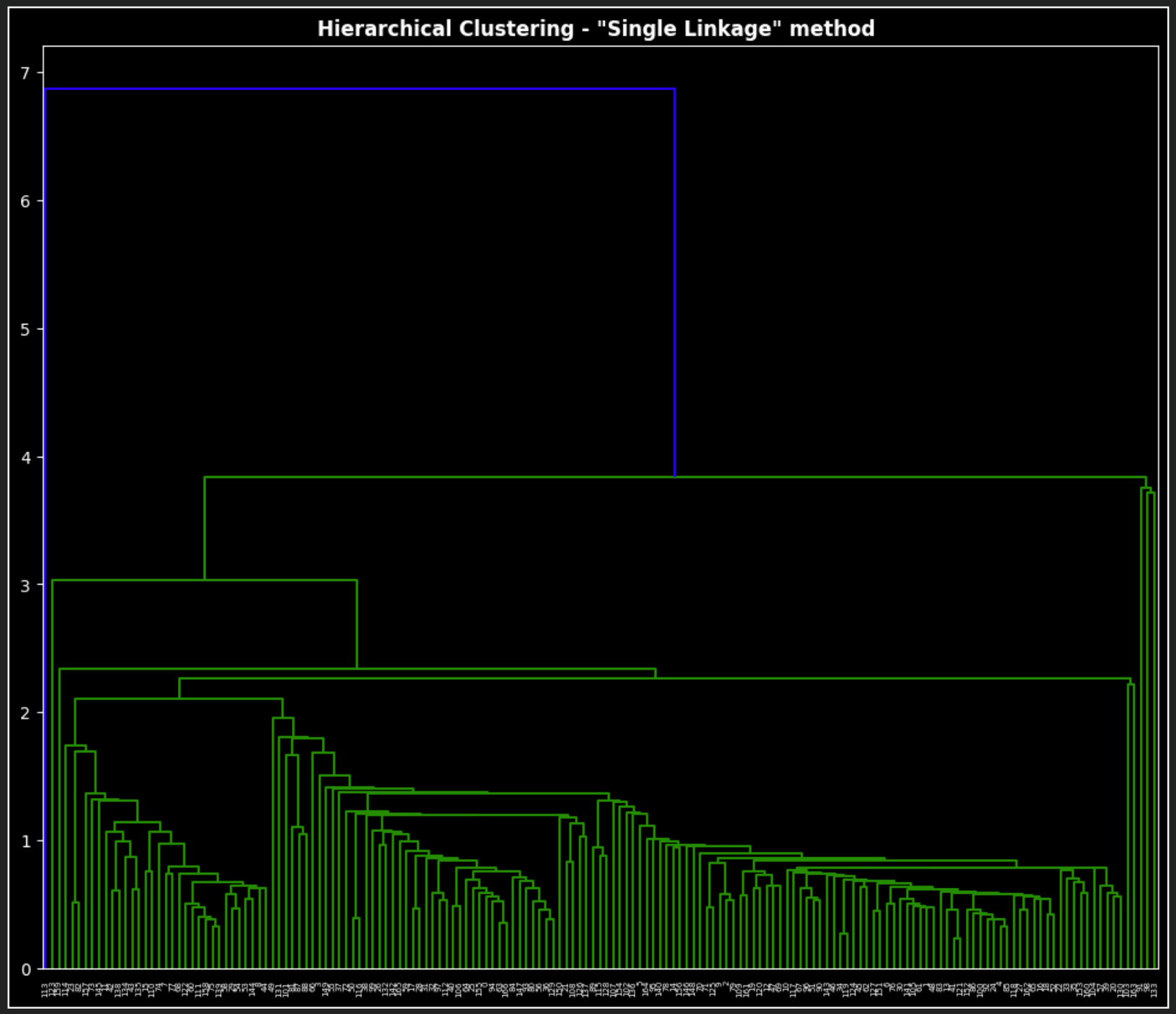


Looking at the above observations:

- SSD maintained steady slope of droppings
- but, for number of cluster = 3, SSD follows further steady path
- Considering final number of clusters are 3 for further analysis.

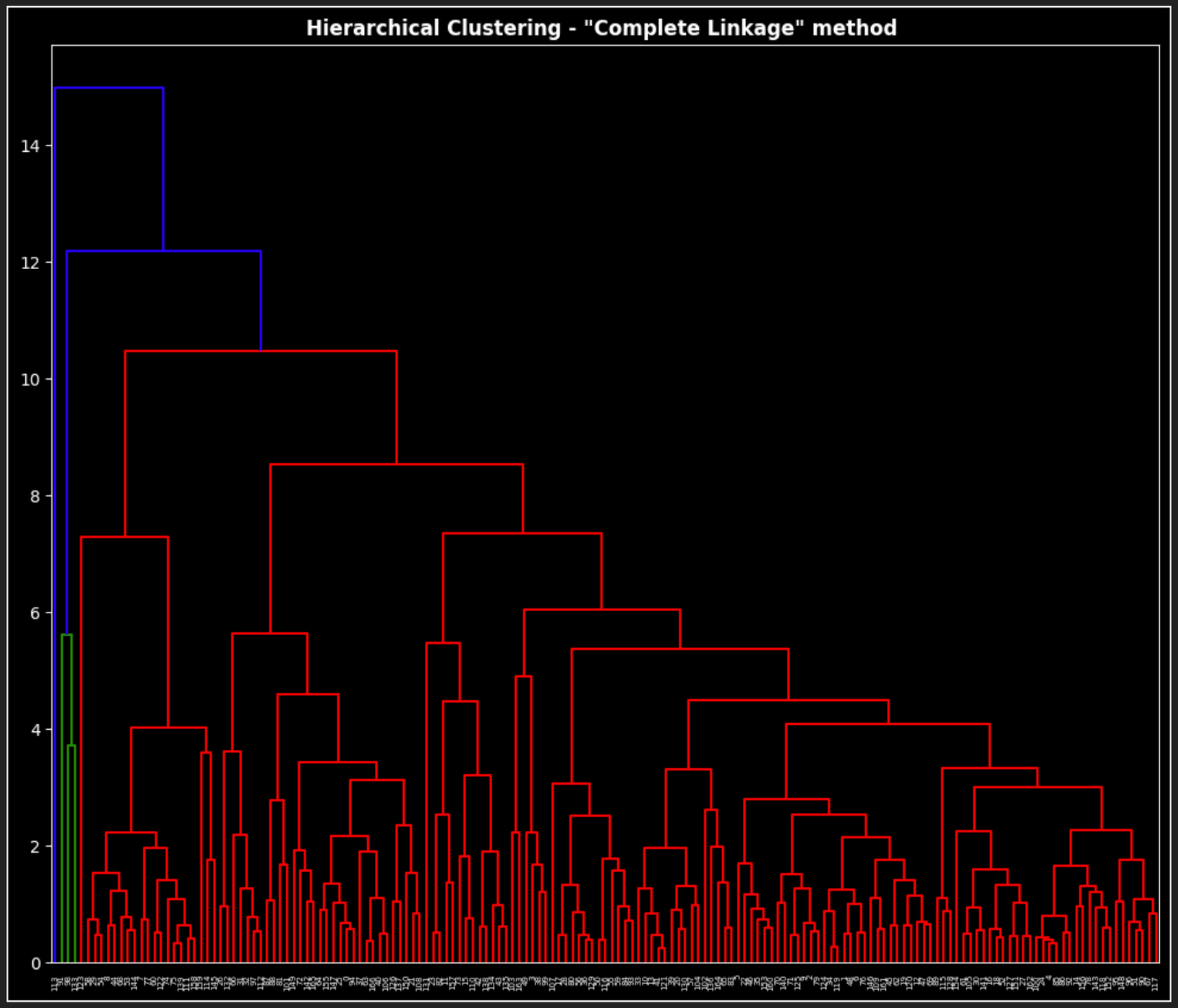
CLUSTERING – HIERARCHICAL METHODOLOGY

Using ‘SINGLE LINKAGE’ method:



CLUSTERING – HIERARCHICAL METHODOLOGY

Using ‘COMPLETE LINKAGE’ method:



With above plotting and confirming 3 clusters would be optimal for the operations

CLUSTER PROFILE – NEW DATA-PLANE PCA DATA

Five Principle Component vectors converted actual data points into 5 independent non-colinear data and final Clustered groups are maintaining a mean of PC1 to PC5 as following

cluster_id_PCA_and_KMeans	PC1	PC2	PC3	PC4	PC5
0	0.175190	-0.138616	0.027608	-0.759568	0.148538
1	2.769799	-0.212660	0.065634	0.868601	-0.170155
2	-2.434654	0.410627	-0.099614	0.692214	-0.135141

CLUSTER PROFILE – ORIGINAL DATA SET

cluster_id_PCA_and_KMeans	gdpp	child_mort	income
0	6486.45	21.93	12305.60
1	42494.44	5.00	45672.22
2	1922.38	92.96	3942.40

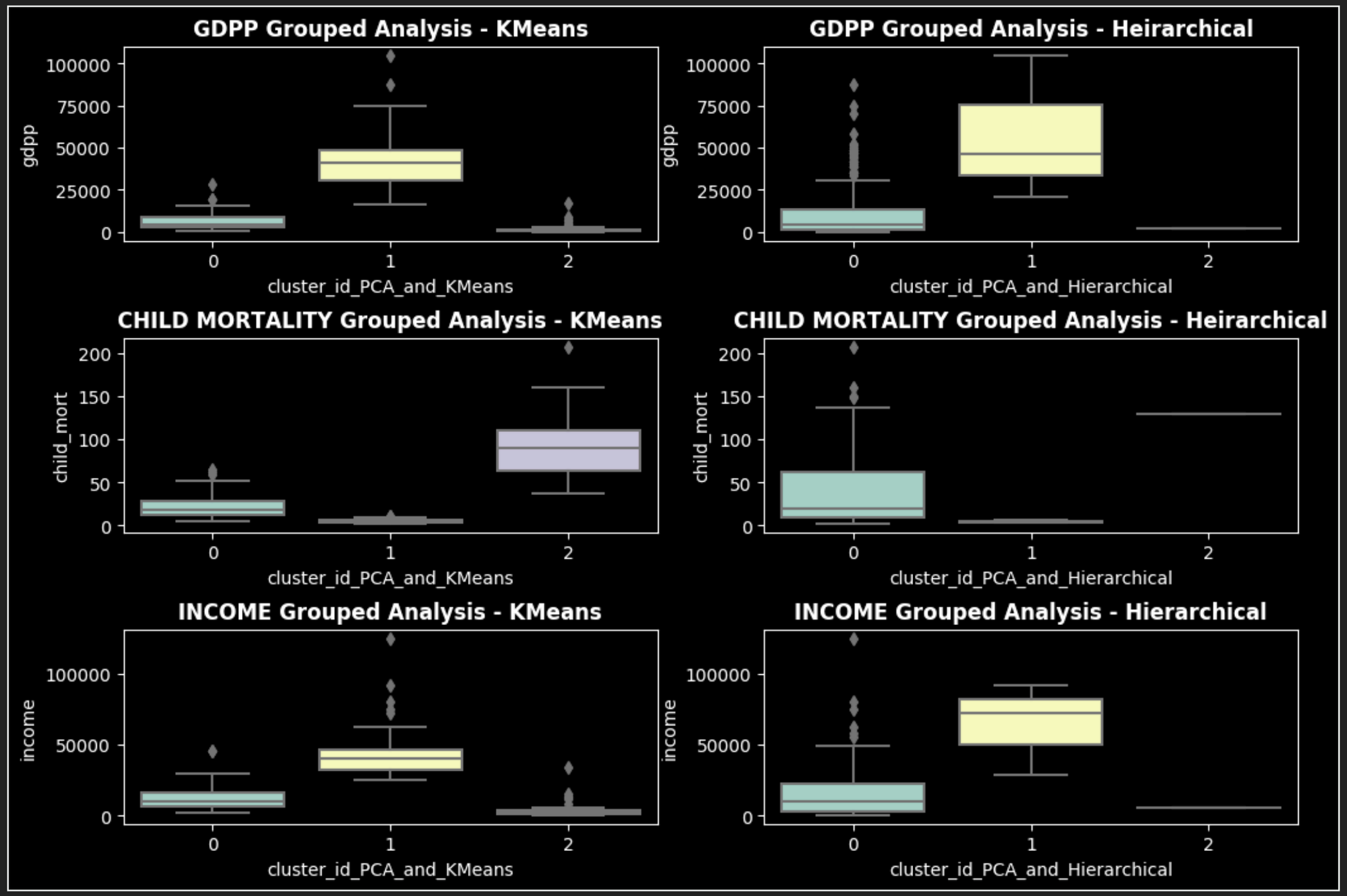
Cluster 0:
- Developing Countries

Cluster 1:
- Developed Countries

Cluster 2:
- Poor developing Countries - Need of help

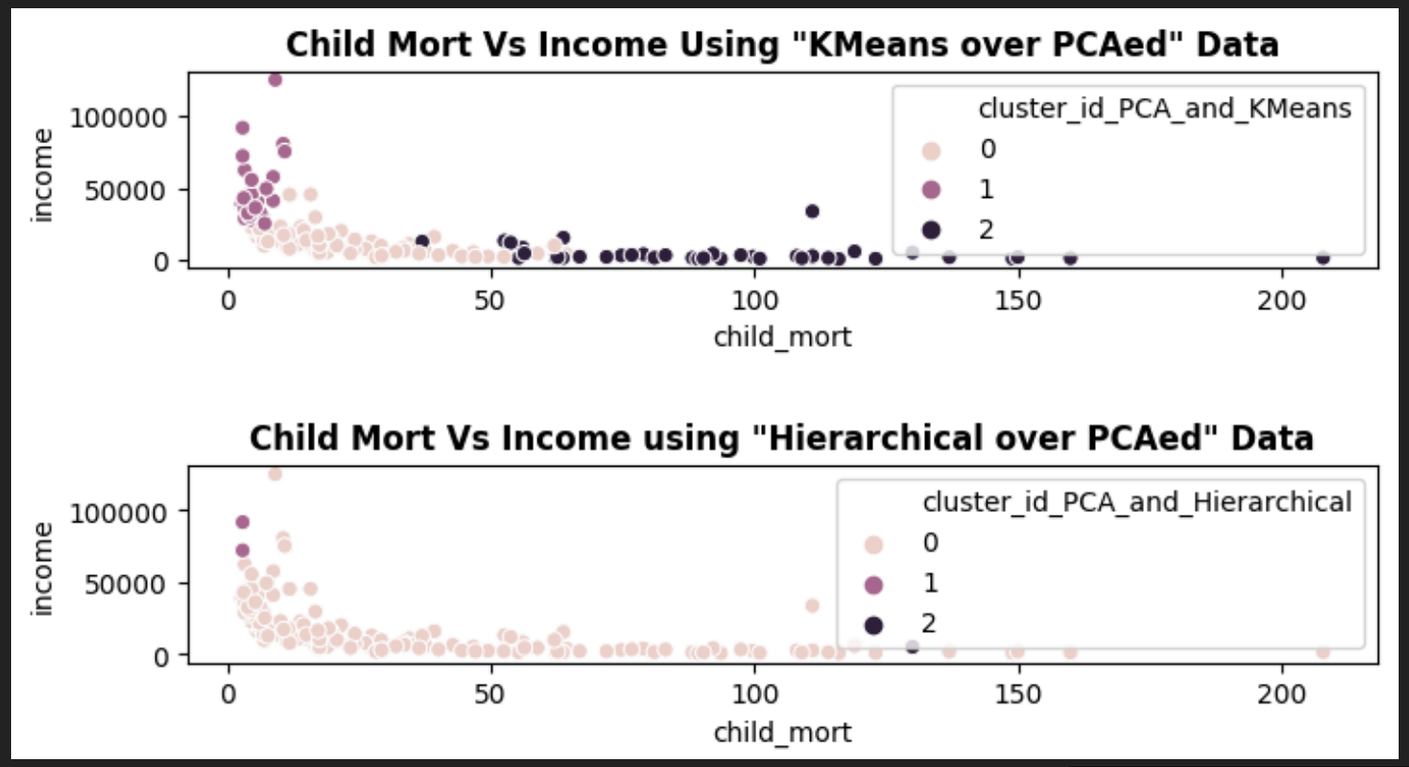
ORIGINAL DATA – CLUSTERING BEHAVIOUR ANALYSIS

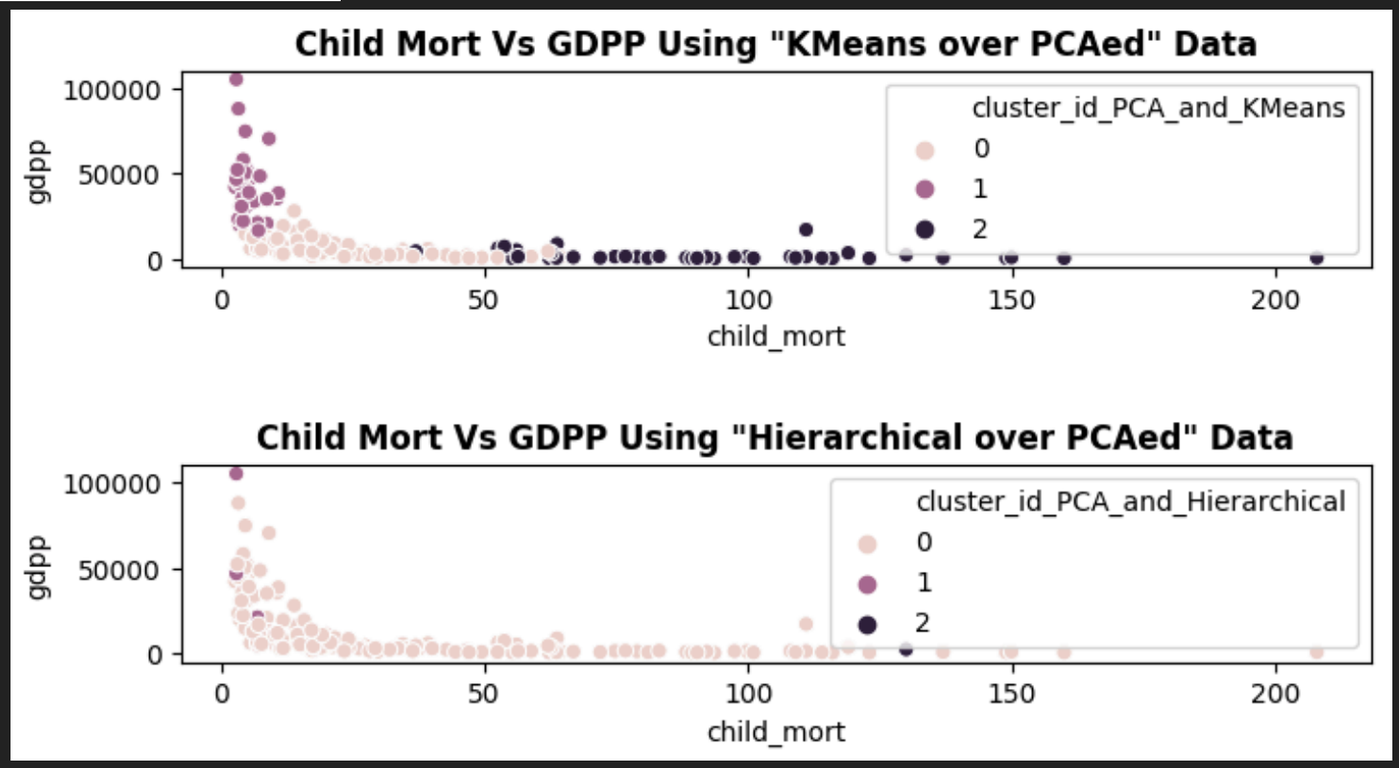
Comparing the clustering behaviour in between KMEANs and HIERARCHICAL on Original Data set.



ORIGINAL DATA – CLUSTERING BEHAVIOUR ANALYSIS

Comparing the clustering behaviour in between KMEANs and HIERARCHICAL on Original Data set.





ORIGINAL DATA – CLUSTERING BEHAVIOUR ANALYSIS,..CONTD

Looking at the KMeans Clustering:

1. Cluster 0:

- These countries are consistently, maintaing average median values of GDPP and INCOME, so does having average median of CHILD_MORTALITY rate too.
- So, safe to consider this as '**DEVELOPING COUNTRIES**'

2. Cluster 1:

- having a median of GDPP at around 4500 and Income median at around close to 50000, we are having Child_mortality very minimum compared to rest of the countries with median almost close to Zero.
- Thus these countries, signifying "**DEVELOPED COUNTRIES**"

3. Cluster 2:

- Having GDPP median at almost close to Zero and so does the income, we have very high Child Mortality rate with a median at around close to 100.
- Thus these countries, falls under "**POOR DEVELOPING COUNTRIES**" ==> which definitely in dire need of HELP

CONCLUSION

List Of Poor Developing Countries:

0	Afghanistan	50	Eritrea	106	Mozambique
3	Angola	55	Gabon	108	Namibia
17	Benin	56	Gambia	**112**	**Niger**
21	Botswana	59	Ghana	113	Nigeria
25	Burkina Faso	63	Guinea	116	Pakistan
26	**Burundi**	64	Guinea-Bissau	126	Rwanda
28	Cameroon	66	Haiti	129	Senegal
31	**Central African Republic**	72	Iraq	132	Sierra Leone
32	Chad	80	Kenya	137	South Africa
36	Comoros	81	Kiribati	142	Sudan
37	**Congo, Dem. Rep.**	84	Lao	147	Tanzania
38	Congo, Rep.	87	Lesotho	149	Timor-Leste
40	Cote d'Ivoire	**88**	**Liberia**	150	Togo
49	Equatorial Guinea	93	Madagascar	155	Uganda
		94	Malawi	165	Yemen
		97	Mali	166	Zambia
		99	Mauritania		

Out of above list, following 5 countries are in direst need of help

37	**Congo, Dem. Rep.**
88	**Liberia**
26	**Burundi**
112	**Niger**
31	**Central African Republic**

These 5 countries are selected for the reasons:

1. Most of the countries with high Child_mortality are due to the less income and top-3 most child nourishment ob served in above 5 list of countries.
2. Less GDPP can definitely impacting the life expectancy, style of human life in that country.