CLUSTERING ASSIGNMENT

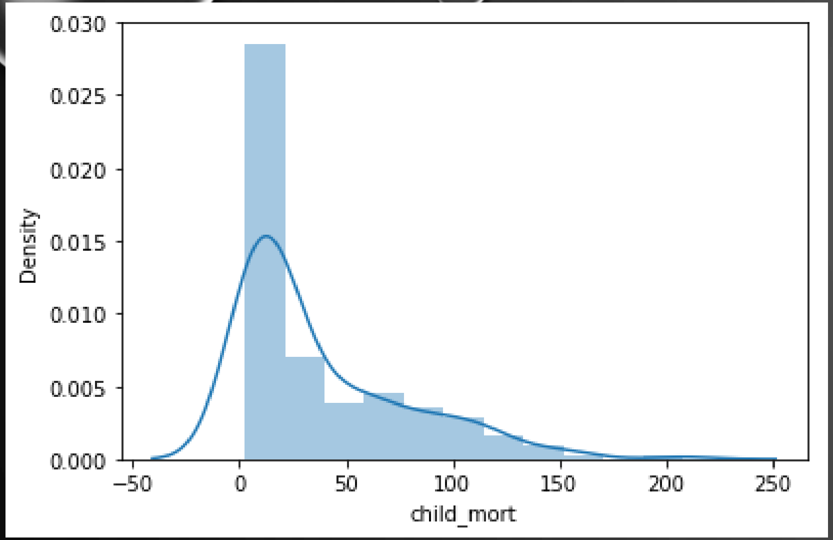
- BY RAMANATHAN RAHUL

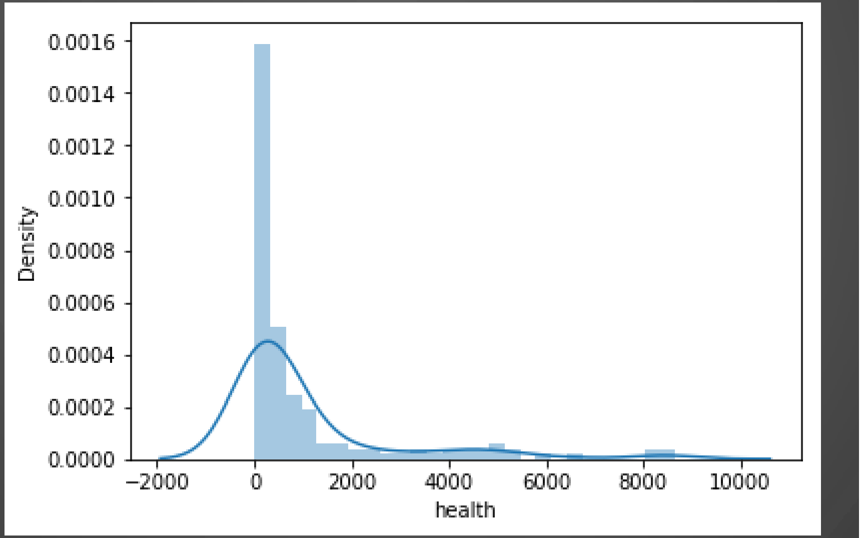
INTRODUCTION

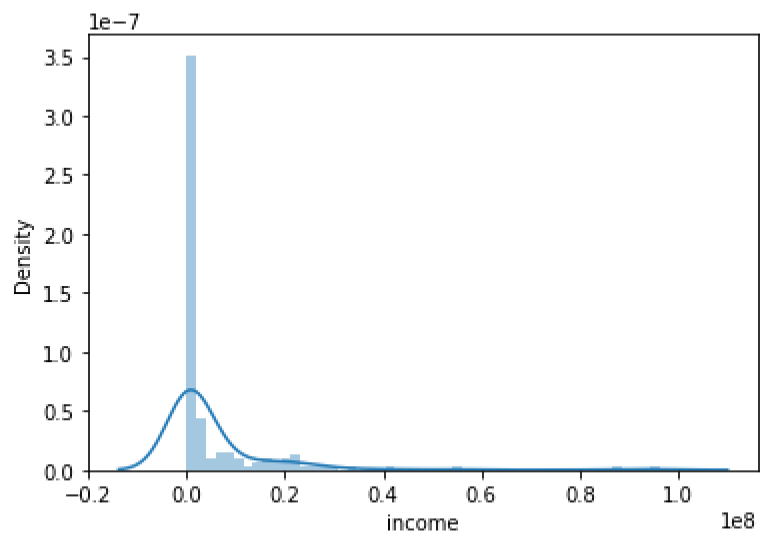
- The goal of this assignment is to help an NGO to categorize the countries in the given dataset using some socio-economic and health factors that determine the overall development of the country.
- The following steps are performed on the given dataset:
  - Data quality check
  - EDA: univariate and bivariate
  - Outlier
  - Hopkin's test
  - Scaling
  - K-means clustering:
    - Find the best value of k using SSD elbow, silhouette score
    - Using the final value of k, perform the kmeans analysis
  - Visualize it using scatterplot
  - Perform cluster profiling: GDPP, CHILD_MORT and INCOME.
  - Hierarchical clustering: single linkage, complete linkage

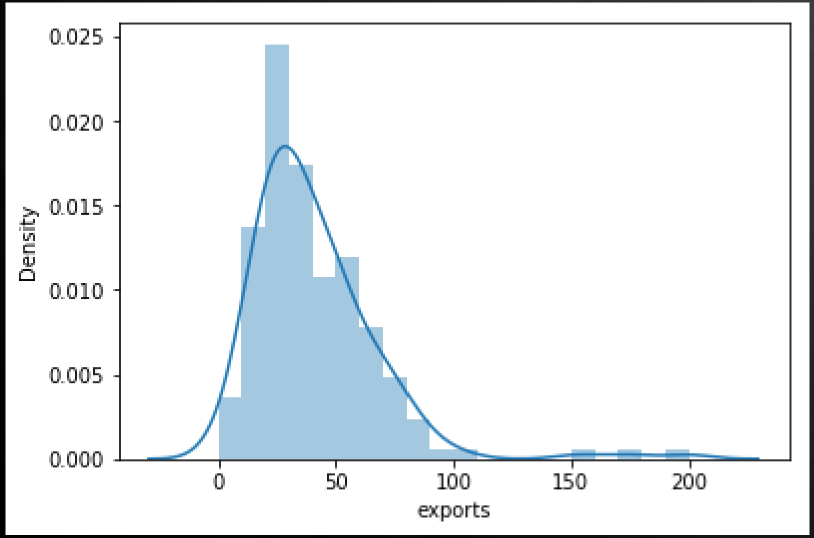
VISUALIZING DISTRIBUTION OF THE DATASET

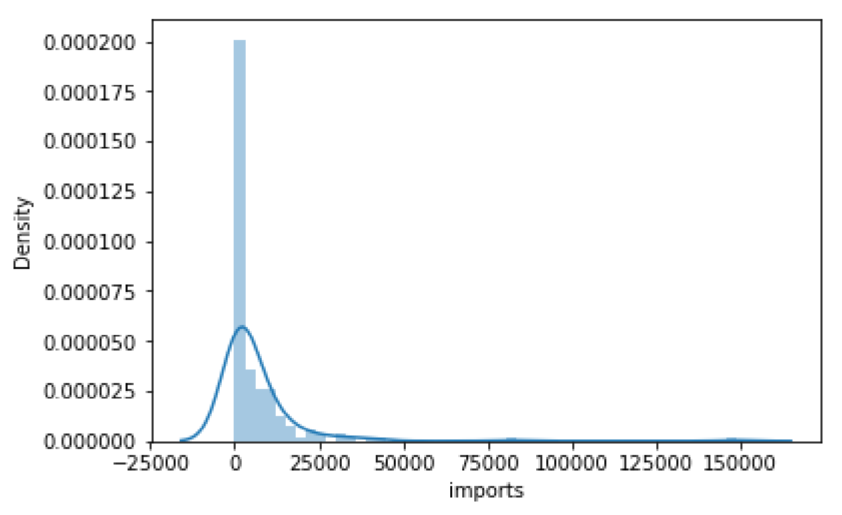
- Inference from the distribution plots (plots in the future slides):
  - The columns 'income', 'health' and 'imports' are normally distributed, we can say that they do not have any internal groupings.
  - The columns 'life_expec', 'gdpp' and 'total_fer'seem to be bimodal distributions, hence there may exist some internal grouping.
  - The remaining columns are not normally distributed as their mean is not at 0, which means there could be skewness in the data or the underlying distribution is non-normal.

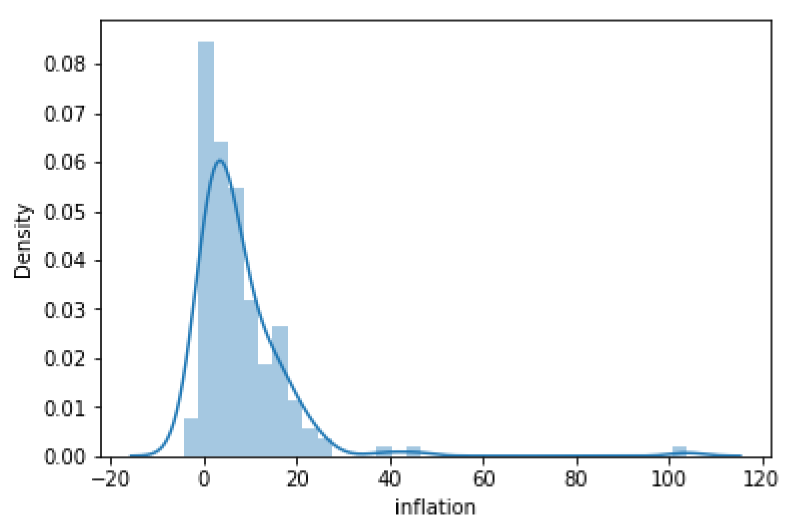

0.030
0.025
0.020
0.015
0.010
0.005
0.000
Density
−50
0
50
100
150
200
250
child_mort

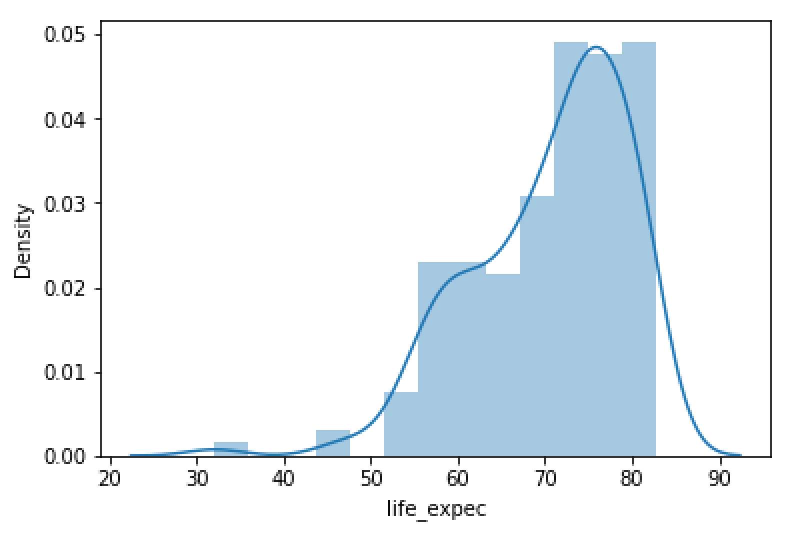

0.0016
0.0014
0.0012
0.0010
0.0008
0.0006
0.0004
0.0002
0.0000
Density
−2000
0
2000
4000
6000
8000
10000
health

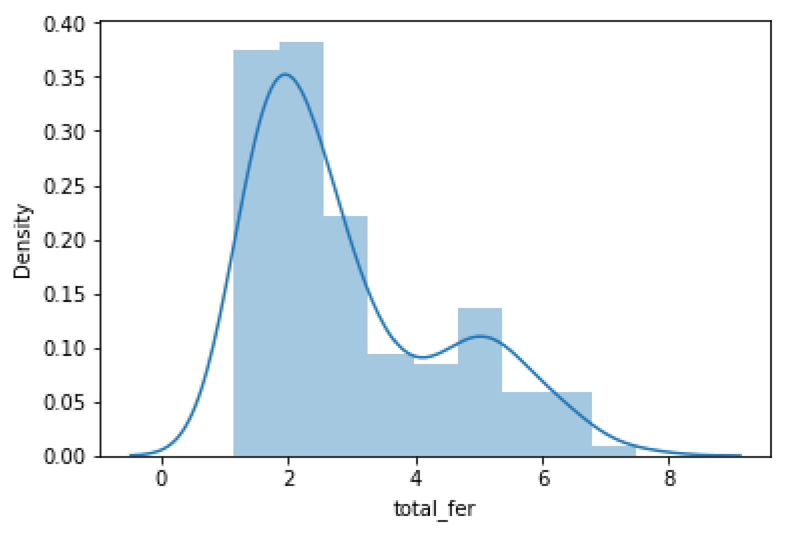

1e−7
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Density
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
income
1e8

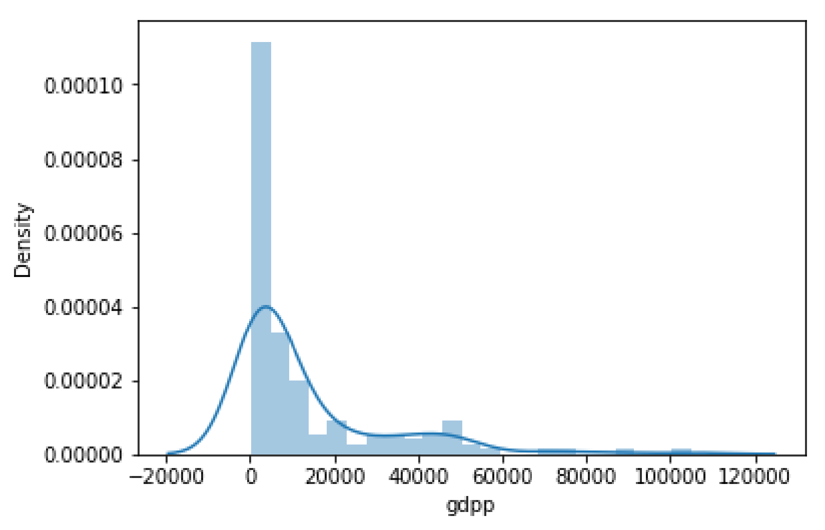

0.025
0.020
0.015
0.010
0.005
0.000
Density
0
50
100
150
200
exports


0.000200
0.000175
0.000150
0.000125
0.000100
0.000075
0.000050
0.000025
0.000000
Density
−25000
0
25000
50000
75000
100000
125000
150000
imports


0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
Density
−20
0
20
40
60
80
100
120
inflation

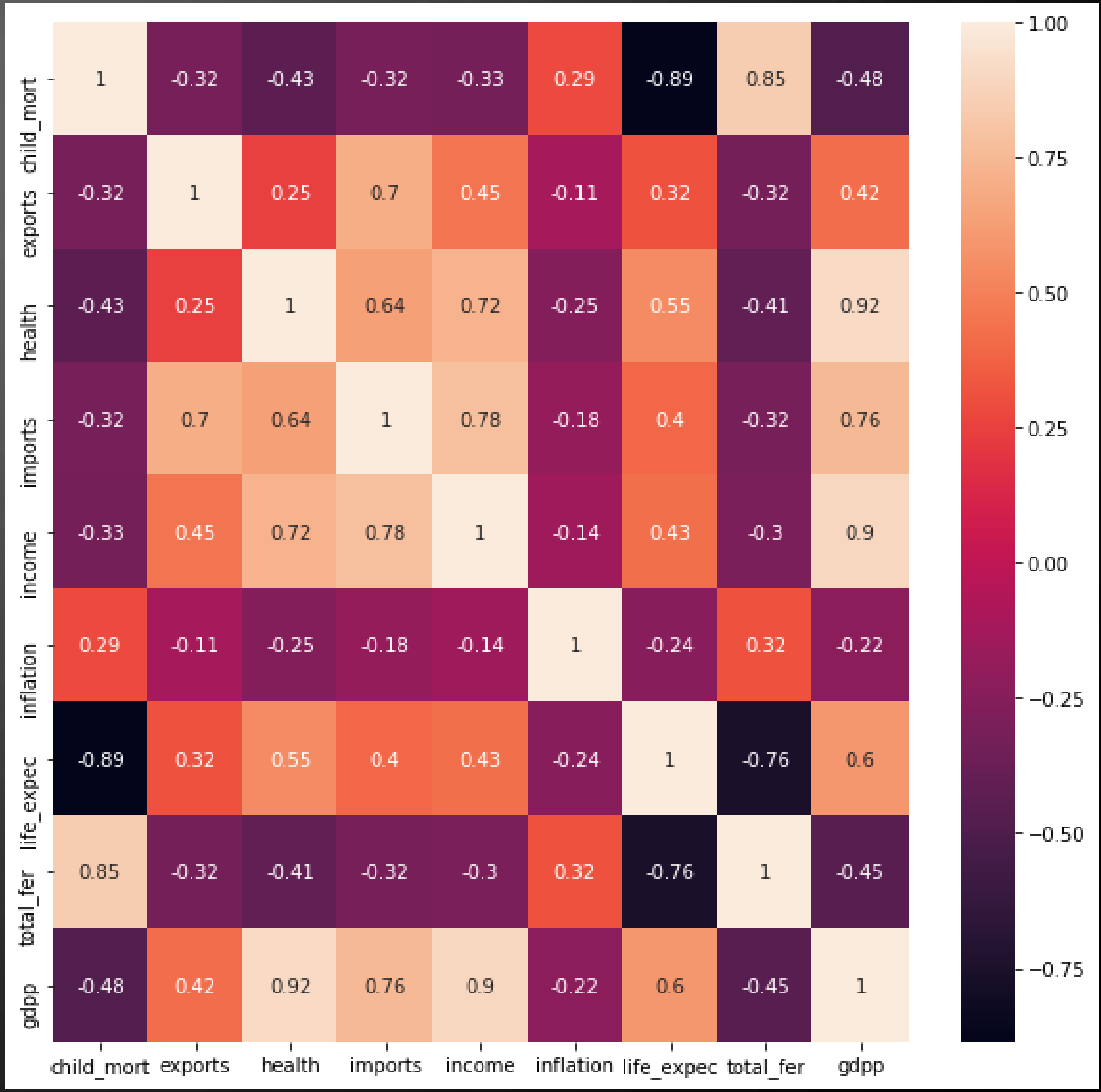

Density
0.05
0.04
0.03
0.02
0.01
0.00
20
30
40
50
60
70
80
90
life_expec


Density
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
2
4
6
8
total_fer


Density
0.00010
0.00008
0.00006
0.00004
0.00002
0.00000
−20000
0
20000
40000
60000
80000
100000
120000
gdpp

VISUALIZING THE CORRELATION OF THE VARIABLES

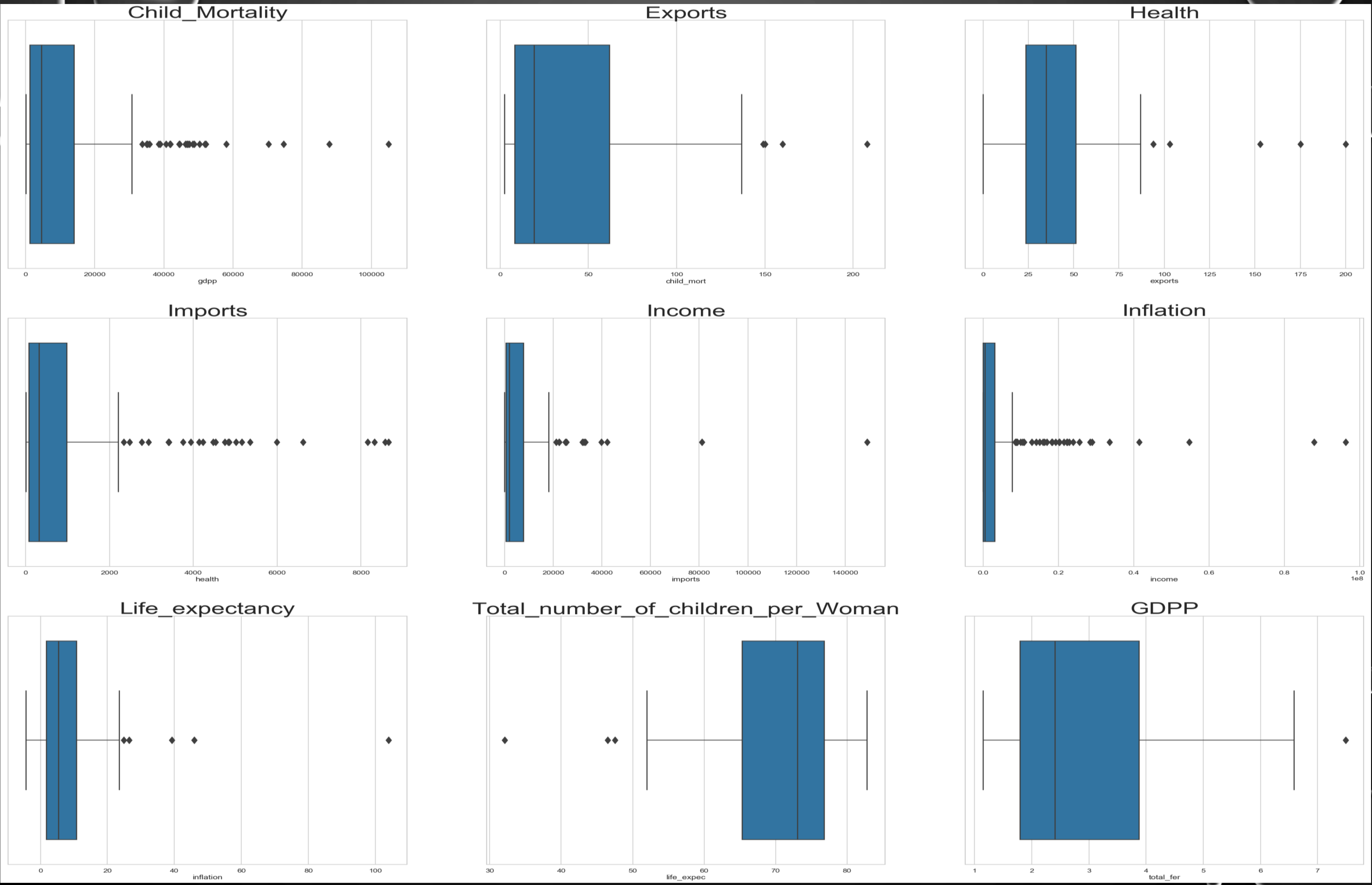
- Inferences from the heatmap:
  - The columns ‘health’, ‘gdpp’, ‘child_mort’, ‘total_fer’ and ‘income’ show high positive correlation.
  - The following set of columns are moderately & positively correlated: 'health’, 'imports', 'income', 'exports’, 'gdpp’ and 'life_expec’.
  - There are columns like 'life_expec', 'child_mort', 'total_fer' having high negative correlation. This shows that the data has multicolinearity.



	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
child_mort	1	-0.32	-0.43	-0.32	-0.33	0.29	-0.89	0.85	-0.48
exports	-0.32	1	0.25	0.7	0.45	-0.11	0.32	-0.32	0.42
health	-0.43	0.25	1	0.64	0.72	-0.25	0.55	-0.41	0.92
imports	-0.32	0.7	0.64	1	0.78	-0.18	0.4	-0.32	0.76
income	-0.33	0.45	0.72	0.78	1	-0.14	0.43	-0.3	0.9
inflation	0.29	-0.11	-0.25	-0.18	-0.14	1	-0.24	0.32	-0.22
life_expec	-0.89	0.32	0.55	0.4	0.43	-0.24	1	-0.76	0.6
total_fer	0.85	-0.32	-0.41	-0.32	-0.3	0.32	-0.76	1	-0.45
gdpp	-0.48	0.42	0.92	0.76	0.9	-0.22	0.6	-0.45	1

BOX PLOTS FOR ANALYZING OUTLIERS

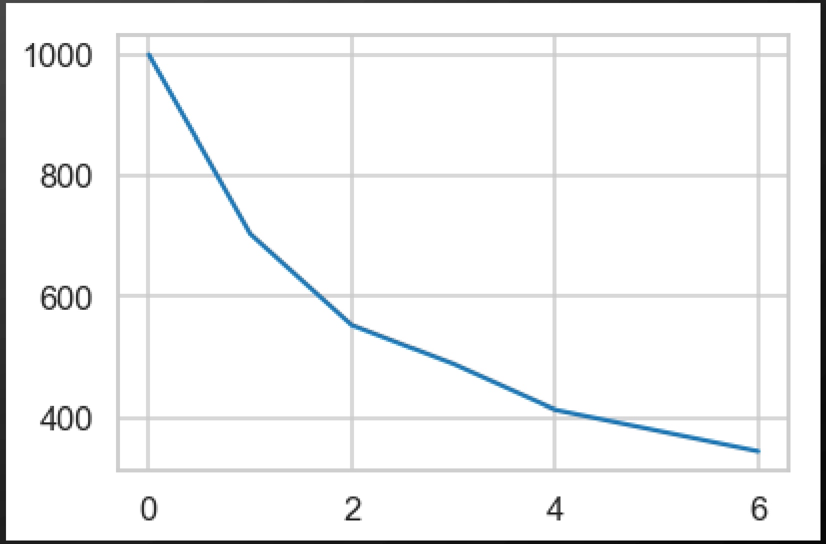
- From the boxplots, we can see that every component has outliers. Now, we need to decide whether to remove them or keep them. We will use the below two approaches:
  - Keep the outliers and do the clustering
  - Remove the outliers and do the clustering


Child_Mortality
0
20000
40000
60000
80000
100000
gdpp
Exports
0
50
100
150
200
child_mort
Health
0
25
50
75
100
125
150
175
200
exports
Imports
0
2000
4000
6000
8000
health
Income
0
20000
40000
60000
80000
100000
120000
140000
imports
Inflation
0.0
0.2
0.4
0.6
0.8
1.0
1e8
income
Life_expectancy
0
20
40
60
80
100
inflation
Total_number_of_children_per_Woman
30
40
50
60
70
80
life_expec
GDPP
1
2
3
4
5
6
7
total_fer

APPROACH 1: K-MEANS CLUSTERING WITH OUTLIERS

K-MEANS ANALYSIS

- Elbow curve using SSD:
  - The elbow plot was plotted for cluster number 2, 3, 4, 5, 6, 7 & 8.
  - From the plot, we can see that the elbow shows at k=3.

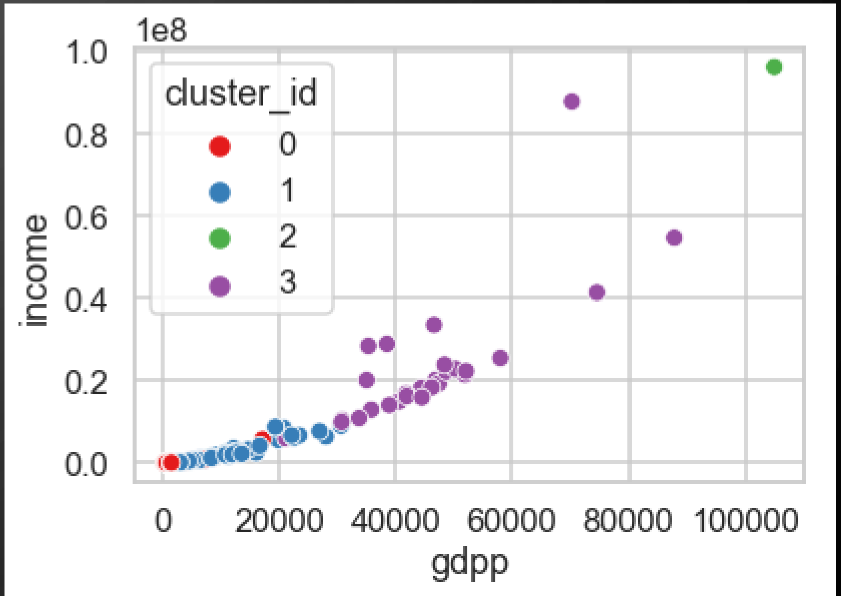


K-MEANS ANALYSIS

- Silhouette analysis:
    - We try to find the optimal k value using the silhouette analysis on the same range of clusters.
    - We get the below values of the silhouette scores:
        - `For n_clusters=2, the silhouette score is 0.423670178948397`
        - `For n_clusters=3, the silhouette score is 0.38878163766741886`
        - `For n_clusters=4, the silhouette score is 0.3916378550223242`
        - `For n_clusters=5, the silhouette score is 0.27146046748399294`
        - `For n_clusters=6, the silhouette score is 0.2809428004371993`
        - `For n_clusters=7, the silhouette score is 0.39012419240069246`
        - `For n_clusters=8, the silhouette score is 0.2695431246442797`
    - We find that, the silhouette score is the highest for k=2, but this k value will not suit our business needs
    - Hence, we proceed with k=4 as it gives precise information and satisfies our business needs.

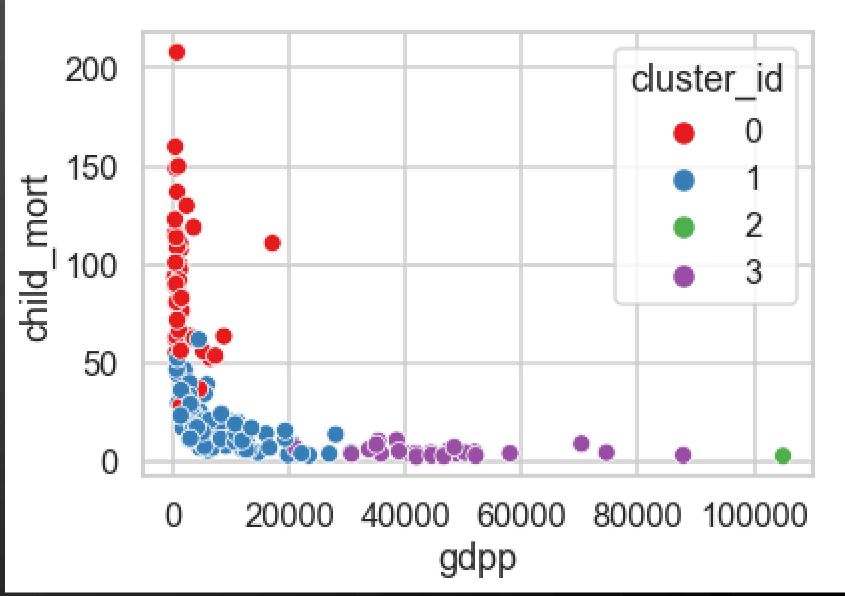
SCATTER PLOT FOR GDPP VS INCOME

- **Inferences from above plot:**
  - Cluster 0 and cluster 1 are of a relatively low gddp and low income group.
  - Cluster 2 and cluster 3 are of high gdpp and higher income groups.



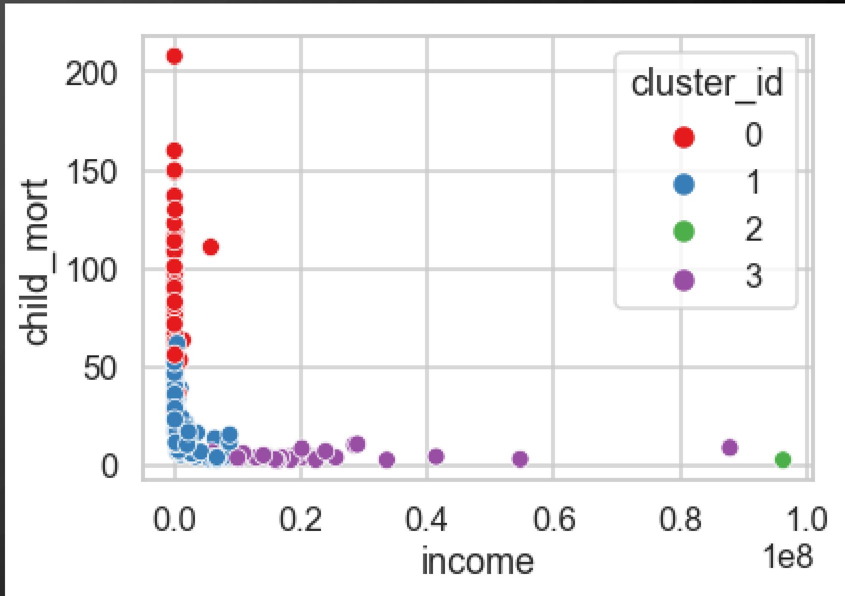
SCATTER PLOT FOR GDPP VS CHILD MORTALITY

- **Inferences from above plot:**
  - Cluster 0 and cluster 1 have high child mortality and a low gdpp values compared to other clusters.
  - Cluster 2 and cluster 3 are of higher gdpp and lower child mortality groups.



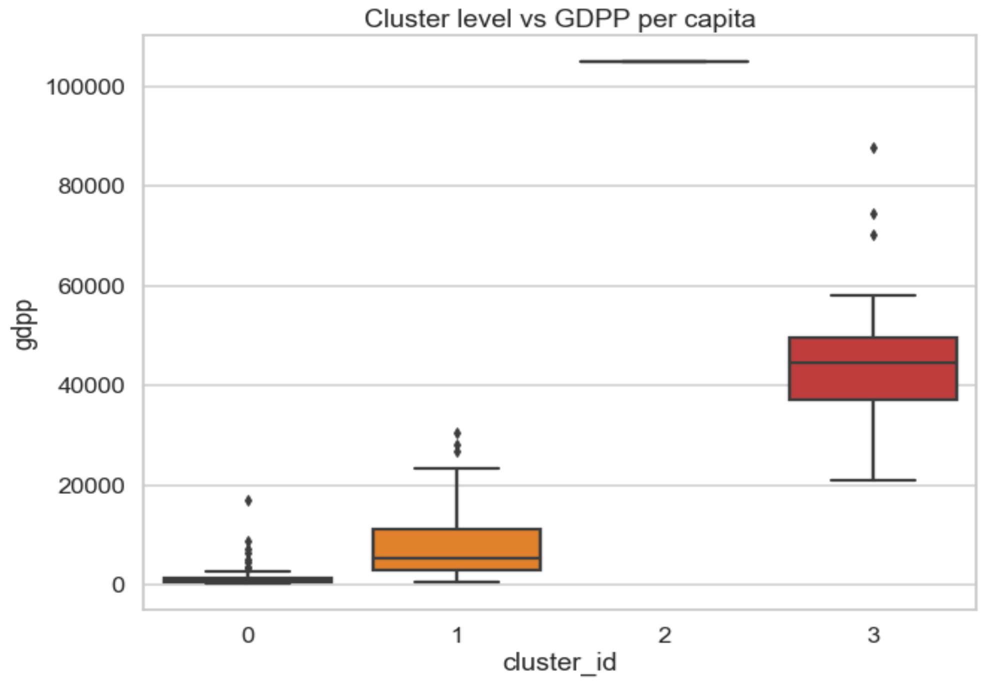
SCATTER PLOT FOR CHILD MORTALITY VS INCOME

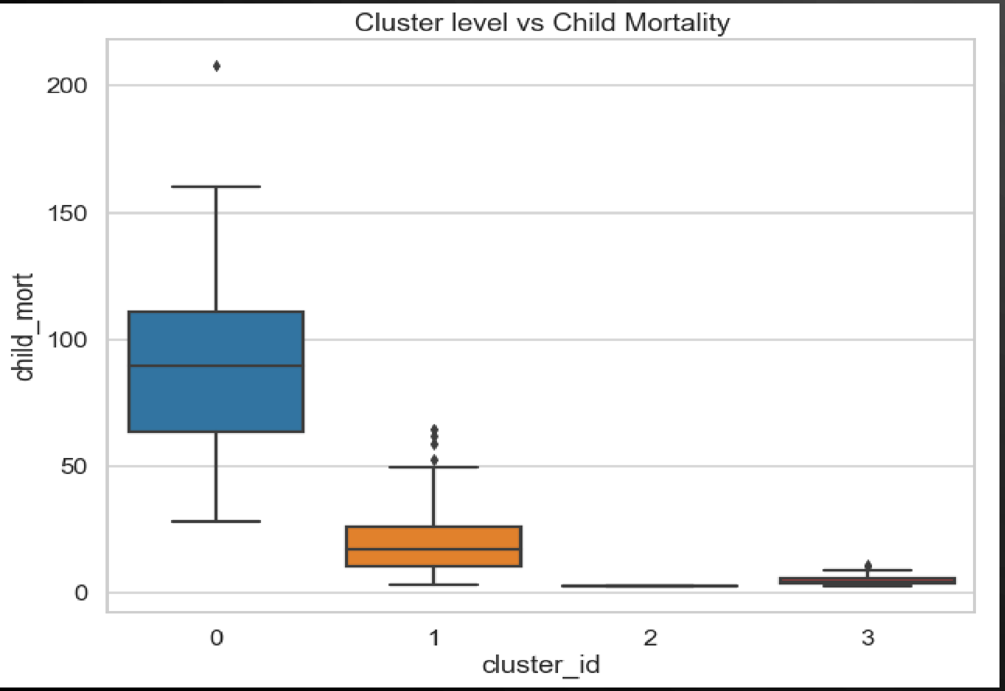
- **Inferences from above plot:**
  - Cluster 0 and cluster 1 have high child mortality and very low income values compared to other clusters.
  - Cluster 2 and cluster 3 are of higher income and much lower child mortality groups.

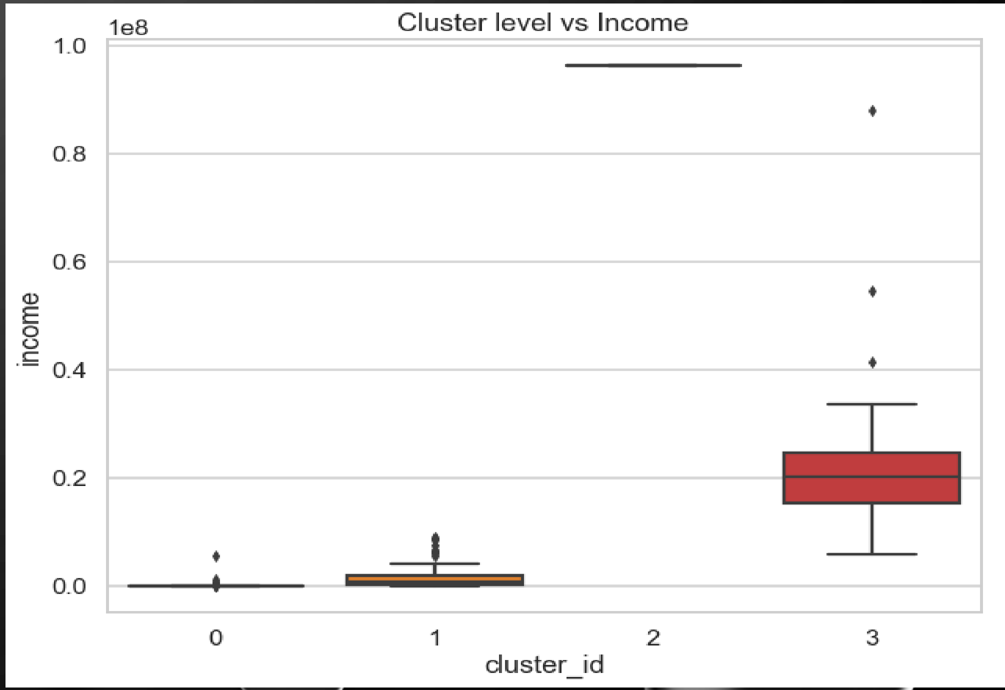


ANALYSIS USING BOXPLOTS

- In summary, we can infer the following points from the boxplots (in next slide):
  - **Cluster label 0:** has the lowest income and lowest gdpp and the highest mortality rate.
  - **Cluster label 1:** has a relatively higher income and gdpp and a lower mortality rate.
  - **Cluster label 2:** has the highest income and gdpp and lowest mortality among the clusters.
  - **Cluster label 3:** has a normal gdpp and income group with a normal child mortality.


Cluster level vs GDPP per capita
gdpp
100000
80000
60000
40000
20000
0
0
1
2
3
cluster_id


Cluster level vs Child Mortality
child_mort
200
150
100
50
0
0
1
2
3
cluster_id


Cluster level vs Income
1e8
income
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
cluster_id

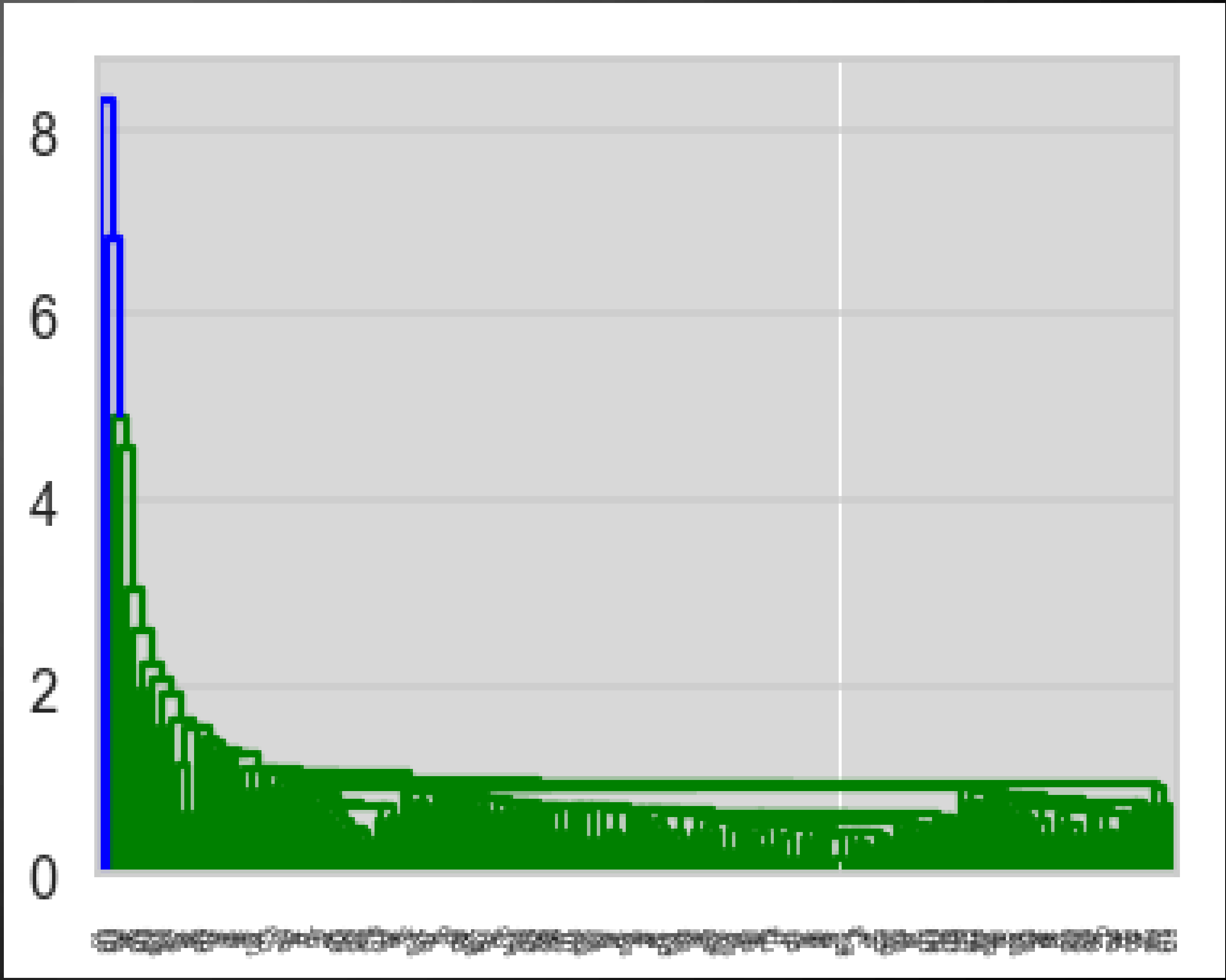
K-MEANS INSIGHTS (WITH OUTLIERS)

- We get the following insights from k-means:
  - There are totally 48 countries from the dataset in need of urgent help as they are having the lowest income, highest child mortality and lowest gdp per capita.
  - Only 1 country are there having good socio-economic and health factors.

APPROACH 1: HIERARCHICAL CLUSTERING WITH OUTLIERS

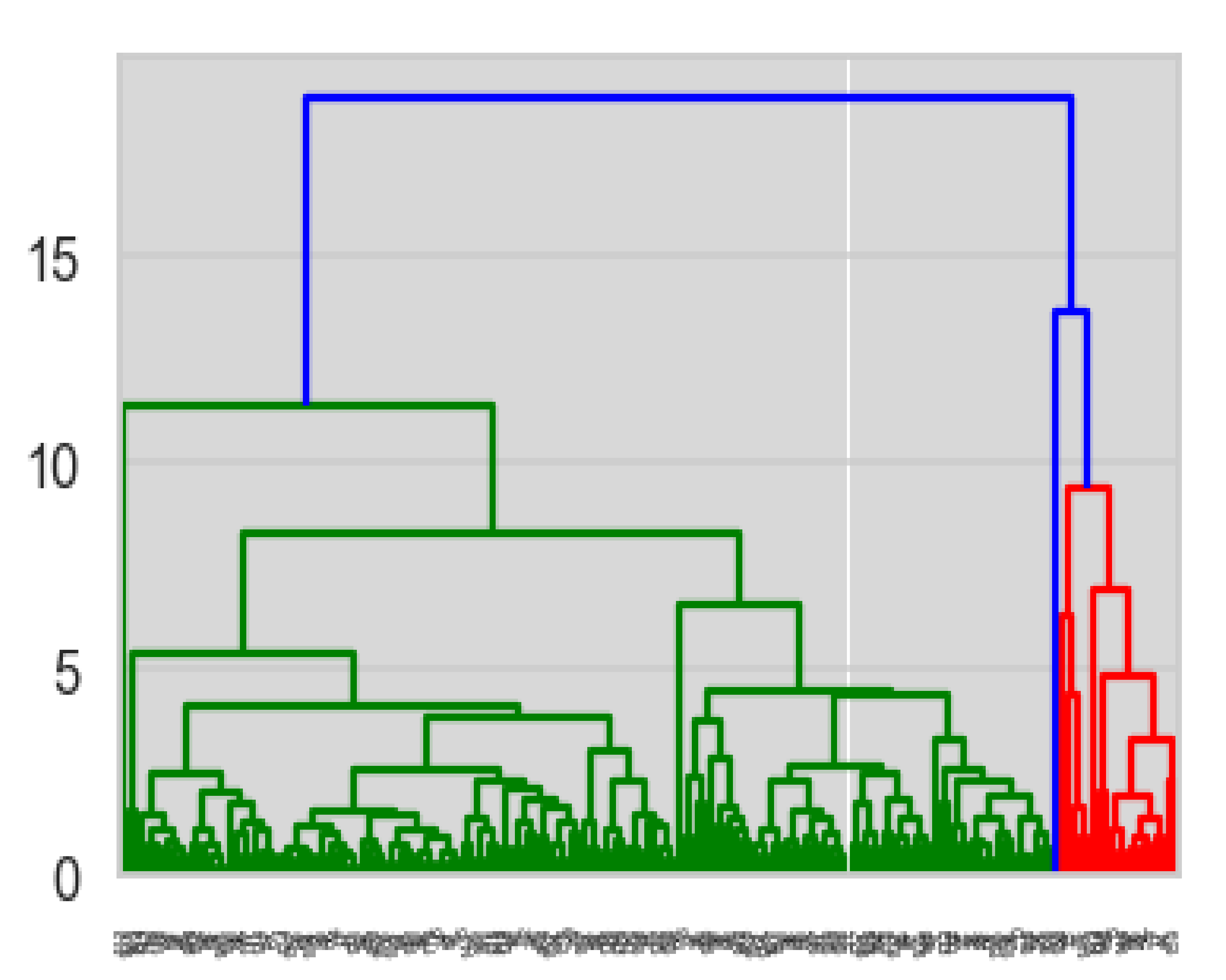
HIERARCHICAL CLUSTERING – SINGLE LINKAGE

- Points to be noted from the dendrogram:
  - This is another method to find the low development countries.
  - As we see from the single linkage dendrogram, it is not clear and does not suit our dataset properly as we can't cut it at a threshold.
  - We will use complete linkage dendrogram for hierarchical clustering.



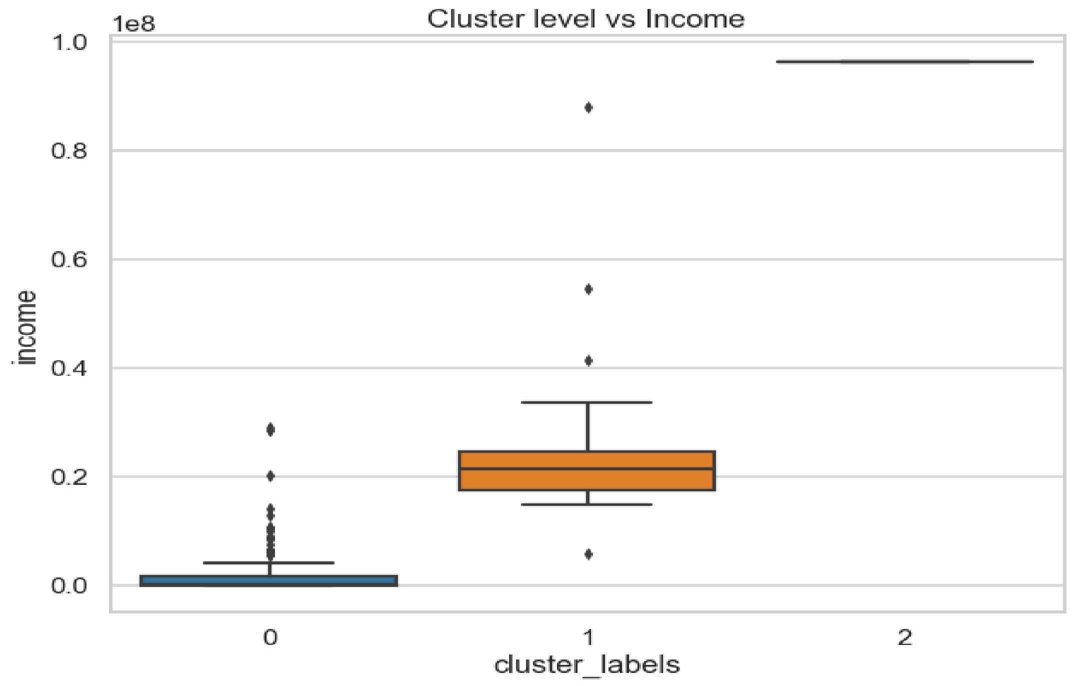
HIERARCHICAL CLUSTERING – COMPLETE LINKAGE

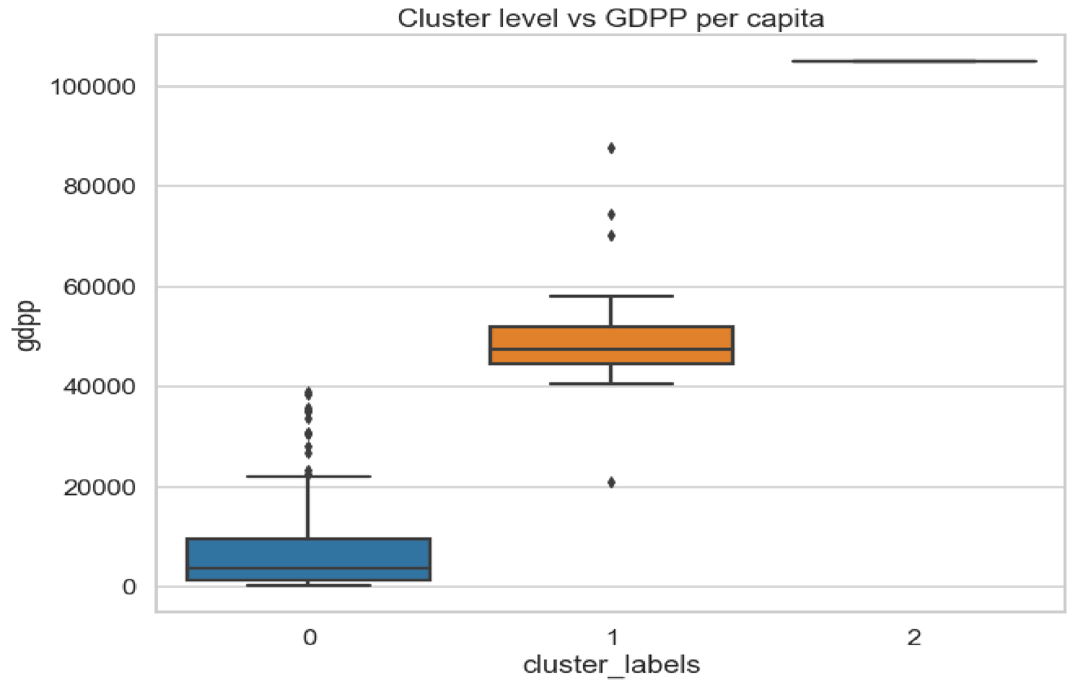
- Points to be noted from the dendrogram:
  - This is a clearer dendrogram plot and it is easier to decide the number of clusters and cut at a threshold value.
  - We will cut at 3 branches, which means we will have 3 clusters.

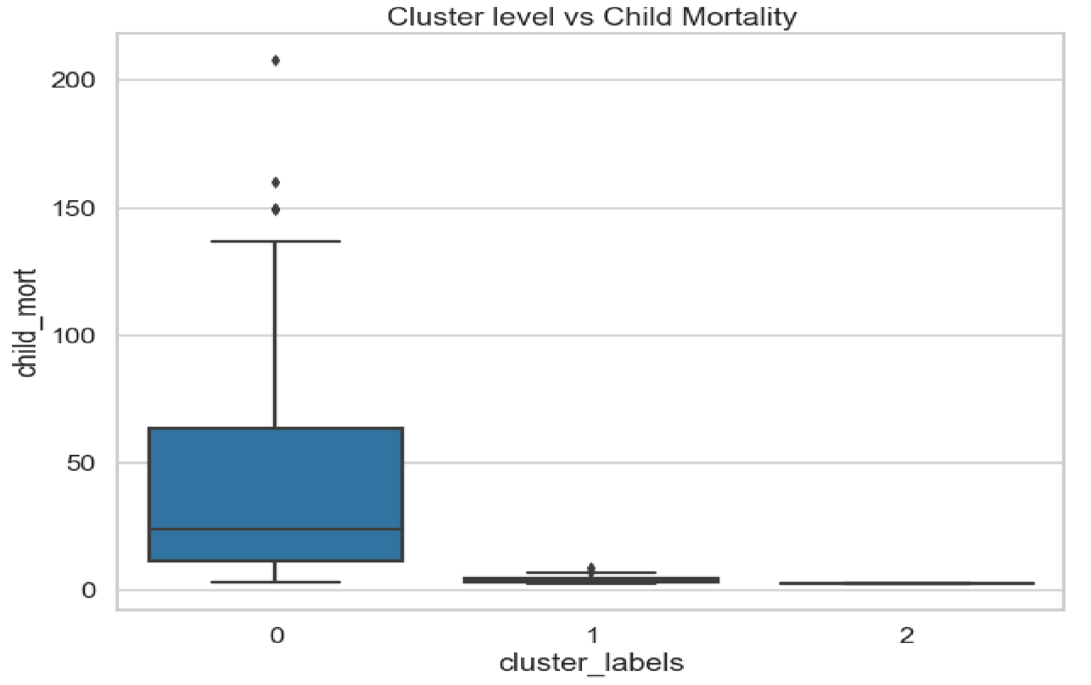


ANALYSIS USING BOXPLOTS

- In summary, we can infer the following points from the boxplots (in next slide):
  - **Cluster 0:** gdpp and income is the lowest compared to other clusters, mortality of children is the highest compared to other clusters.
  - **Cluster 1:** behaves normally in all departments (income, gdpp and children mortality) except for some outliers.
  - **Cluster 2:** gdpp and income is highest compared to other clusters, mortality of children is the least compared to other clusters.


Cluster level vs Income


Cluster level vs GDPP per capita


Cluster level vs Child Mortality

HIERARCHICAL CLUSTERING INSIGHTS (WITH OUTLIERS)

- We get the following insights from hierarchical clustering:
  - There are totally 147 countries from the dataset in need of urgent help as they are having the lowest income, highest child mortality and lowest gdp per capita.
  - Only 1 country are there having good socio-economic and health factors.

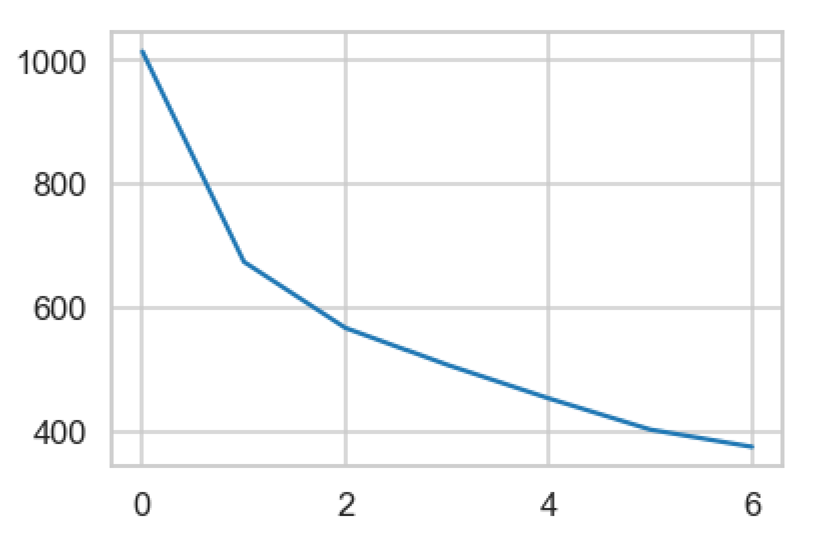
CONCLUSION – WITH OUTLIERS

- **K-means clustering :**
  - Countries that are direst need of aid:
    - Total 48 countries are in this category
  - Countries that are having good socio-economic and health factors:
    - Only 1 country is in this category - luxembourg
- **Hierarchical clustering:**
  - Countries that are direst need of aid
    - Total 147 countries are in this category
  - Countries that are having good socio-economic and health factors
    - Only 1 country is in this category - luxembourg
  - We have seen from both k-means and hierarchical clustering methods that hierarchical clustering selects an extra 99 countries. I would choose the final countries from k-means clustering as it gave a more accurate output compared to hierarchical clustering. I have compared the clusters and visualized from both methods and k-means gave more precise information than hierarchical clustering.

APPROACH 2: K-MEANS CLUSTERING WITHOUT OUTLIERS

K-MEANS ANALYSIS

- Elbow curve using SSD:
  - The elbow plot was plotted for cluster number 2, 3, 4, 5, 6, 7 & 8.
  - From the plot, we can see that the elbow shows at k=3.

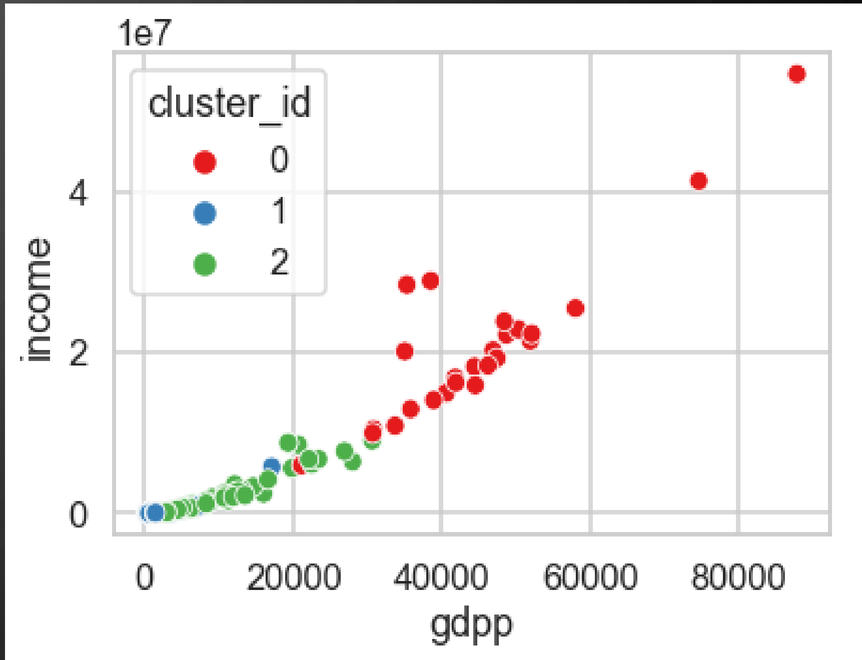


K-MEANS ANALYSIS

- Silhouette analysis:
  - We try to find the optimal k value using the silhouette analysis on the same range of clusters.
  - We get the below values of the silhouette scores:
    - `For n_clusters=2, the silhouette score is 0.5297678399035708`
    - `For n_clusters=3, the silhouette score is 0.4005075430106541`
    - `For n_clusters=4, the silhouette score is 0.34931356620635323`
    - `For n_clusters=5, the silhouette score is 0.3398456050978124`
    - `For n_clusters=6, the silhouette score is 0.29495161627549393`
    - `For n_clusters=7, the silhouette score is 0.28346480847515404`
    - `For n_clusters=8, the silhouette score is 0.24371397409320172`
  - We find that, the silhouette score is the highest for k=2, but this k value will not suit our business needs
  - Hence, we proceed with k=3 as it gives precise information and satisfies our business needs.

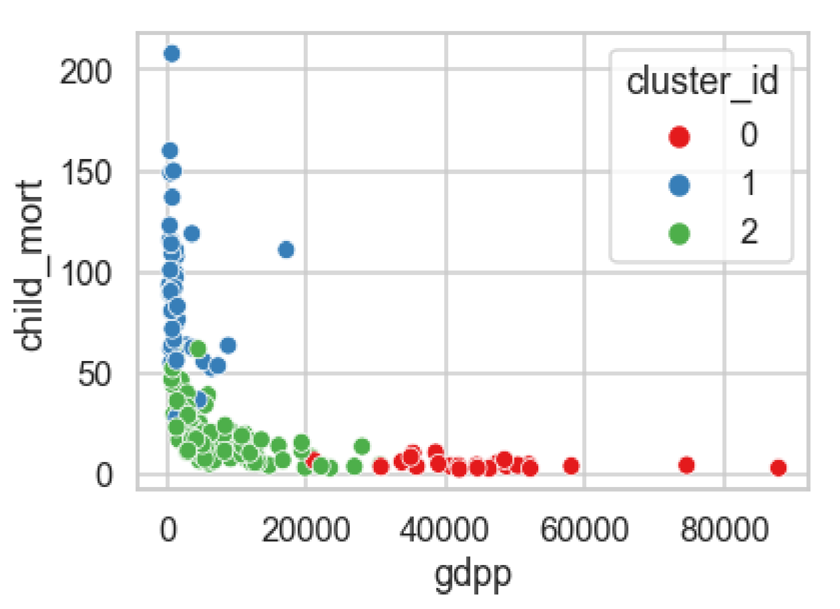
SCATTER PLOT FOR GDPP VS INCOME

- **Inferences from above plot:**
  - Cluster 1 is a low income and low gdpp group whereas cluster 2 is a mixture of both low and high gdpp and income.
  - Cluster 0 is a higher income and gdpp group, higher than clusters 1 and 2.



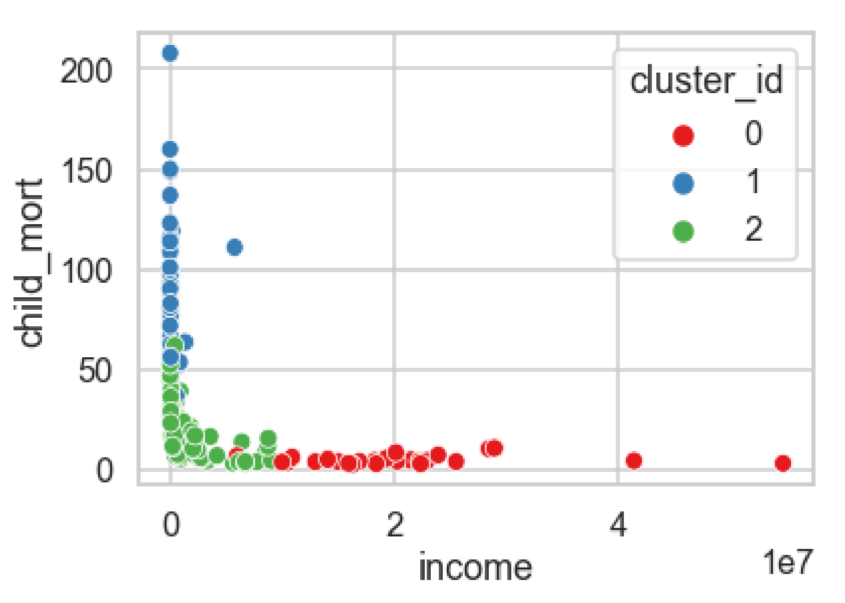
SCATTER PLOT FOR GDPP VS CHILD MORTALITY

- **Inferences from above plot:**
  - Cluster 1 is a high child mortality and low gdpp group whereas cluster 2 is a mixture of relatively lower child mortality and higher gdpp.
  - Cluster 0 is a higher gdpp and lower mortality group, the gdpp being higher than clusters 1 and 2 and mortality being lower than clusters 1 and 2.



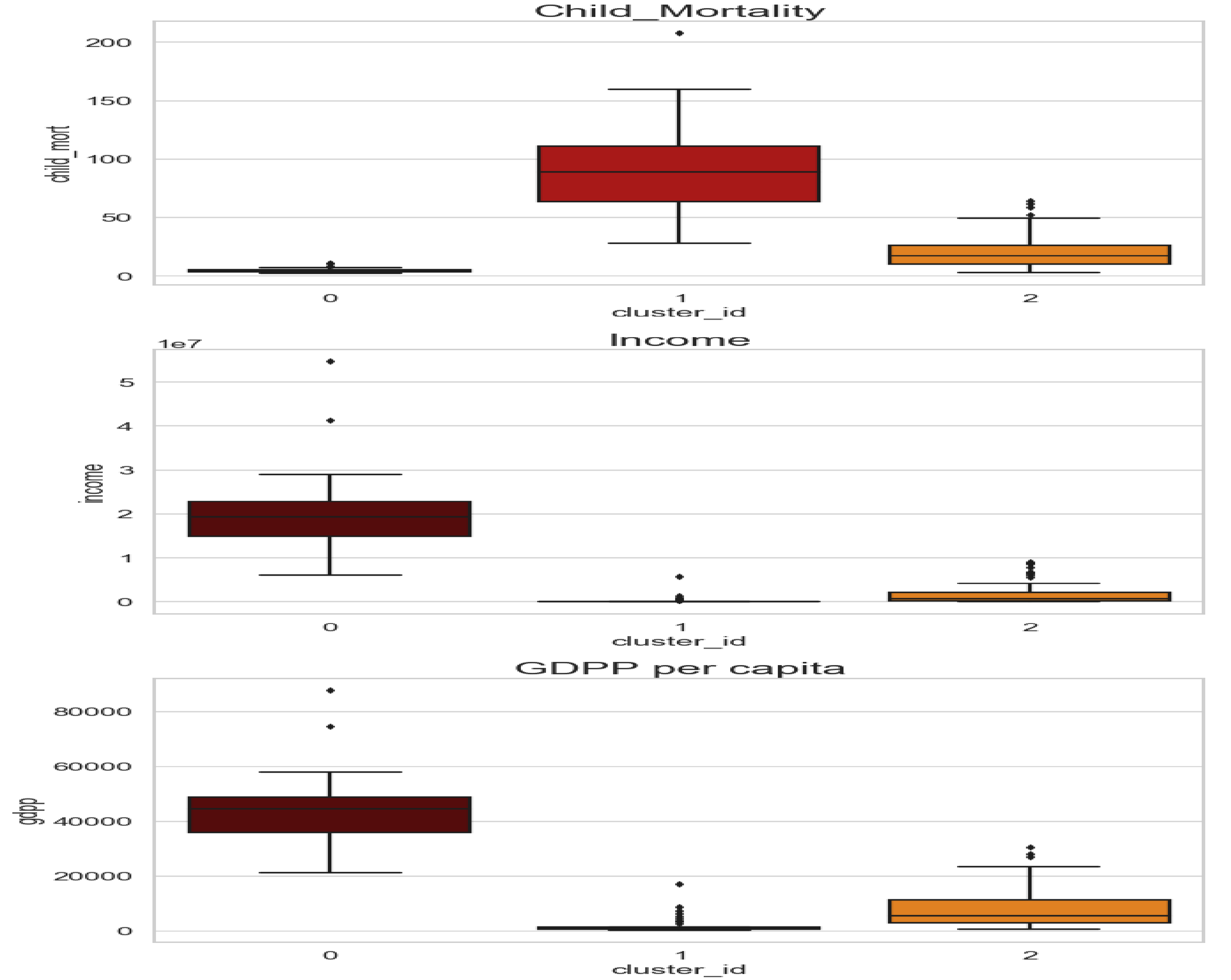
SCATTER PLOT FOR CHILD MORTALITY VS INCOME

- **Inferences from above plot:**
  - Cluster 1 is a high child mortality and low income group whereas cluster 2 is a mixture of relatively lower child mortality and higher income.
  - Cluster 0 is a higher income and lower mortality group, with the gdpp being higher than clusters 1 and 2 and mortality being lower than clusters 1 and 2 in some instances.



ANALYSIS USING BOXPLOTS

- In summary, we can infer the following points from the boxplots (in next slide):
  - **Cluster 0:** gdpp and income is the highest compared to other clusters, mortality of children is the least compared to other clusters.
  - **Cluster 1:** has the highest mortality rate in comparison to other clusters.
  - **Cluster 2:** gdpp, mortality and income is the normal compared to other clusters.


Child_Mortality
child_mort
200
150
100
50
0
0
1
2
cluster_id
Income
1e7
income
5
4
3
2
1
0
0
1
2
cluster_id
GDPP per capita
gdpp
80000
60000
40000
20000
0
0
1
2
cluster_id

K-MEANS INSIGHTS (WITHOUT OUTLIERS)

- We get the following insights from k-means:

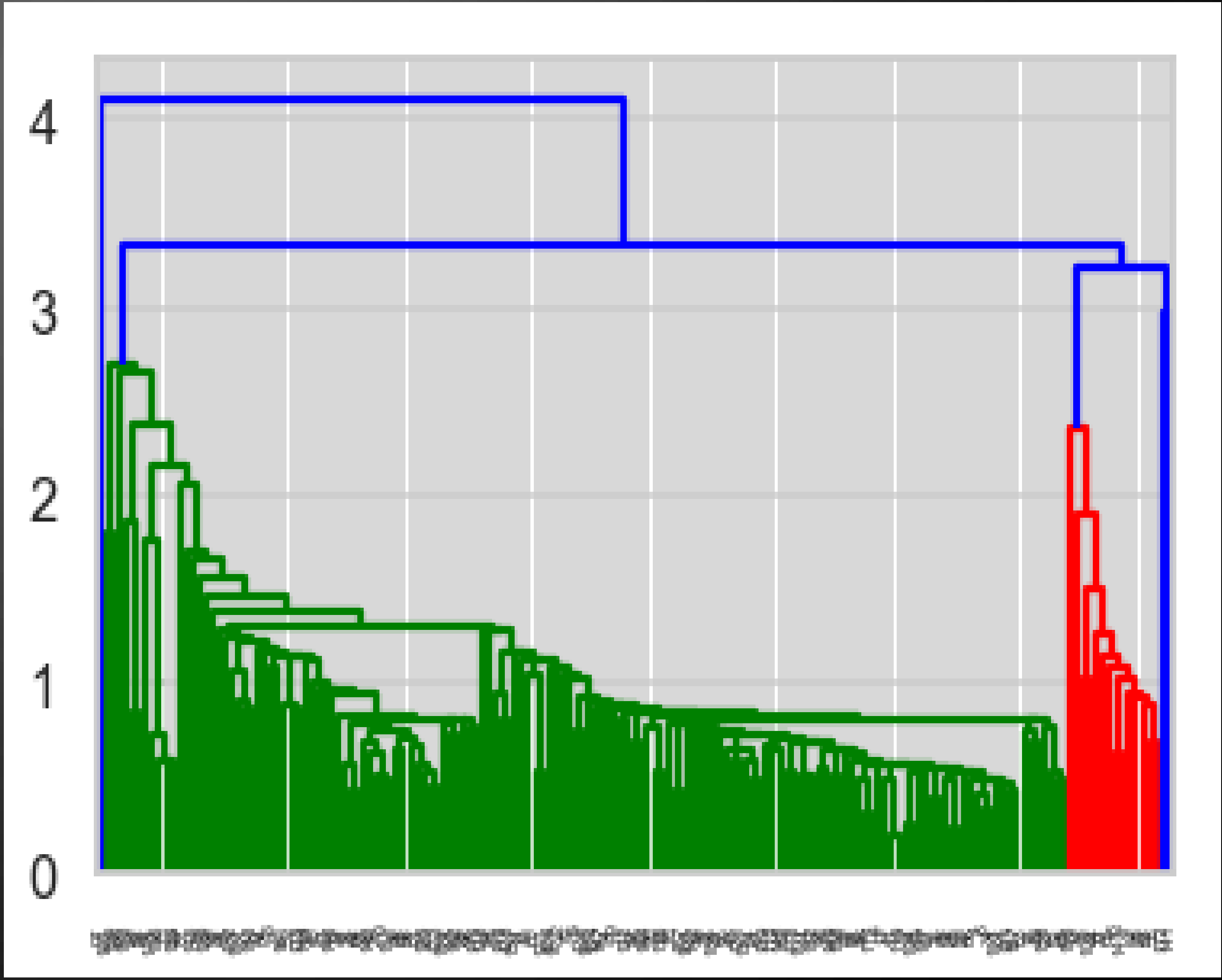
  - There are totally 47 countries from the dataset in need of urgent help as they are having the lowest income, highest child mortality and lowest gdp per capita.

  - There are totally 25 countries having good socio-economic and health factors and hence do not require any help.

APPROACH 2: HIERARCHICAL CLUSTERING WITHOUT OUTLIERS

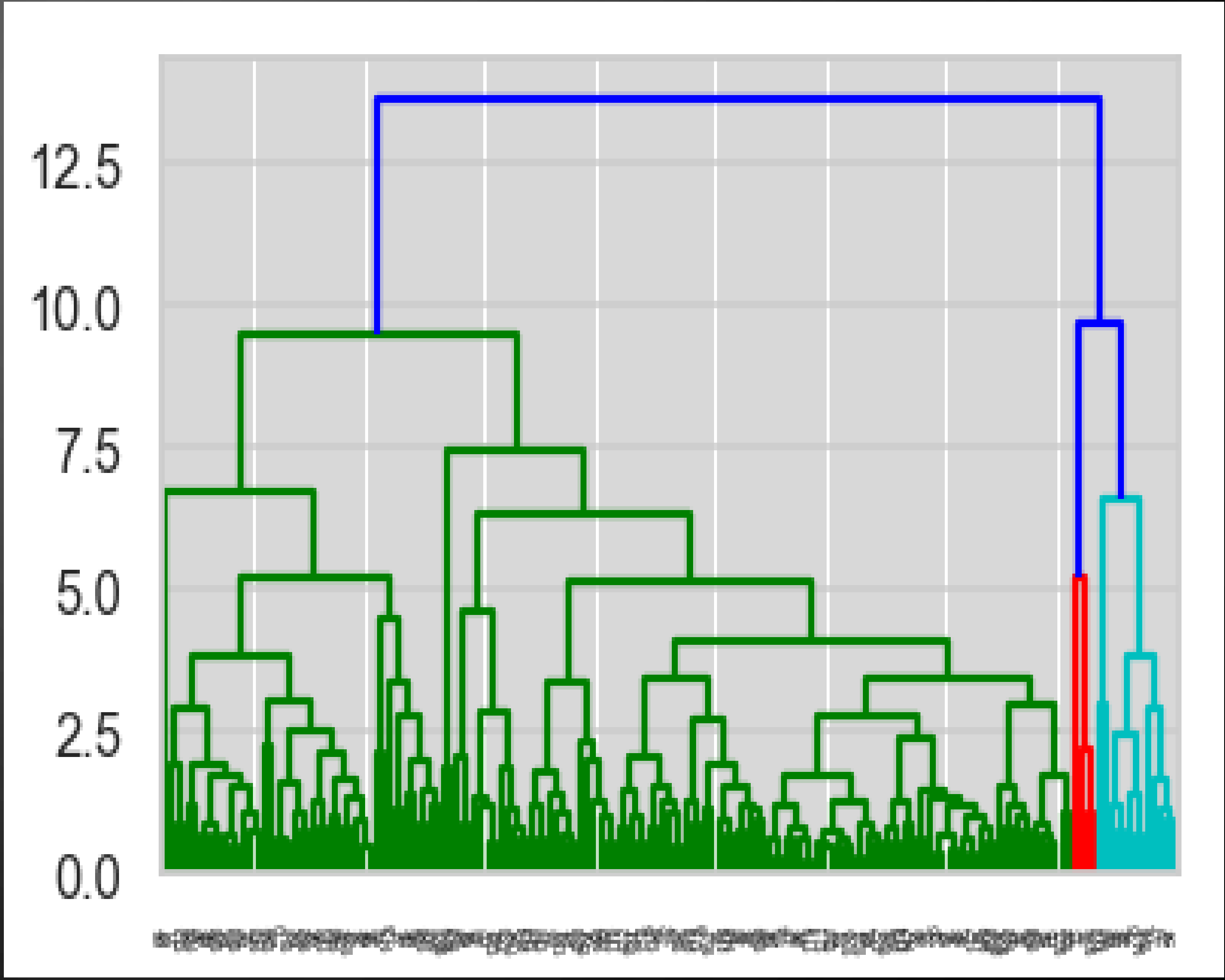
HIERARCHICAL CLUSTERING – SINGLE LINKAGE

- Points to be noted from the dendrogram:
  - This is another method to find the low development countries.
  - As we see from the single linkage dendrogram, it is not clear and does not suit our dataset properly as we can't cut it at a threshold.
  - We will use complete linkage dendrogram for hierarchical clustering.



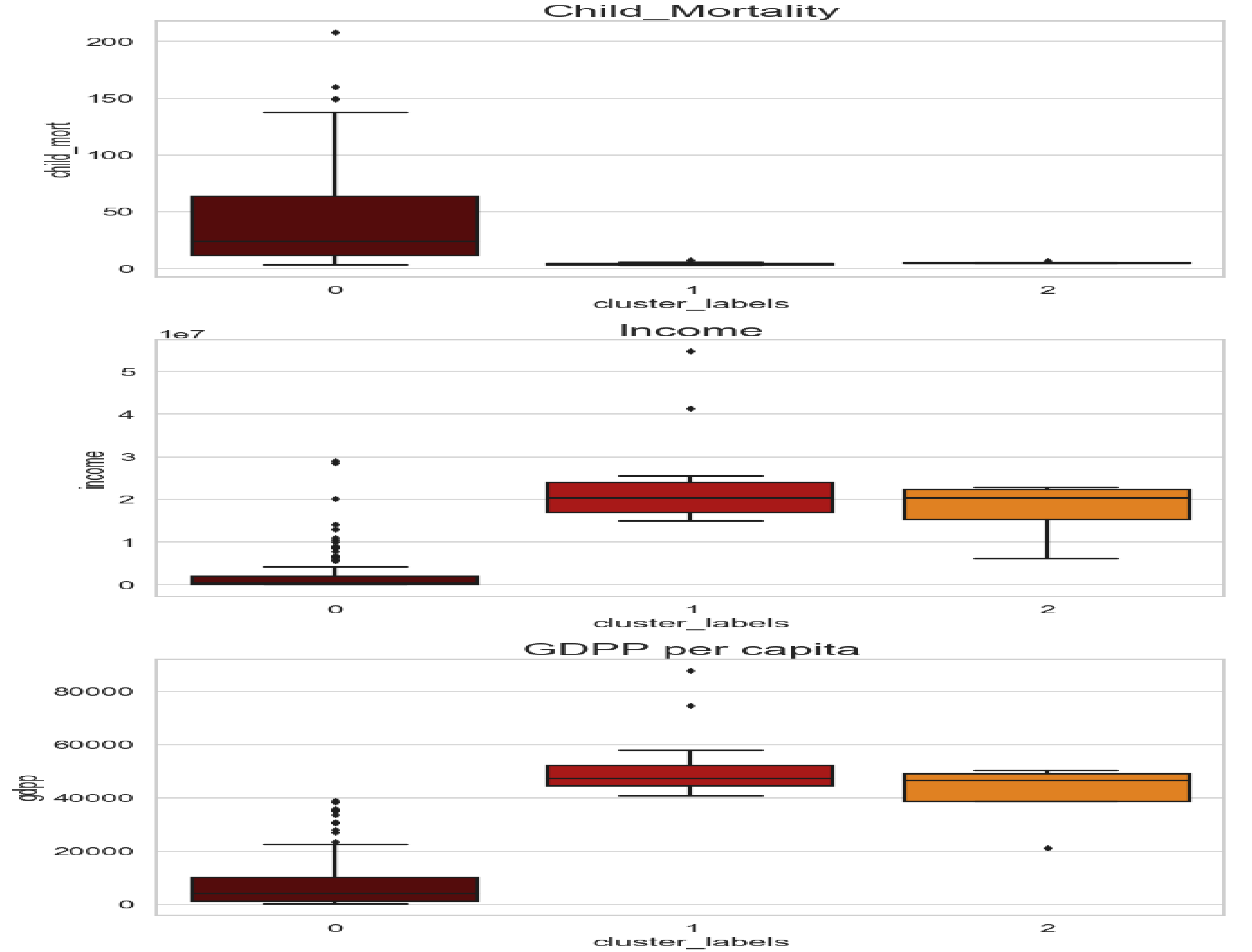
HIERARCHICAL CLUSTERING – COMPLETE LINKAGE

- Points to be noted from the dendrogram:
  - This is a clearer dendrogram plot and it is easier to decide the number of clusters and cut at a threshold value.
  - We will cut at 3 branches, which means we will have 3 clusters.



ANALYSIS USING BOXPLOTS

- In summary, we can infer the following points from the boxplots (in next slide):
  - **Cluster 0:** gdpp and income is the lowest than others clusters, mortality of children is very high compared to other clusters.
  - **Cluster 1:** gdpp and income are having decently high values, mortality of children is very low in here.
  - **Cluster 2:** gdpp and income is slightly lower than clusters 2, mortality of children is the least compared to other clusters.


Child_Mortality
child_mort
200
150
100
50
0
0
1
2
cluster_labels
Income
1e7
income
5
4
3
2
1
0
0
1
2
cluster_labels
GDPP per capita
gdpp
80000
60000
40000
20000
0
0
1
2
cluster_labels

HIERARCHICAL CLUSTERING INSIGHTS (WITHOUT OUTLIERS)

- We get the following insights from hierarchical clustering:
  - There are totally 146 countries from the dataset in need of urgent help as they are having the lowest income, highest child mortality and lowest gdp per capita.
  - There are 4 countries which are having good social-economic and health factors and hence do not require any aid.

CONCLUSION – WITHOUT OUTLIERS

- **K-means clustering:**
  - Countries that are direst need of aid:
    - Total 45 countries are in this category
  - Countries that are having good socio-economic and health factors:
    - Total 25 countries are in this category
- **Hierarchical clustering:**
  - Countries that are direst need of aid
    - Total 147 countries are in this category
  - Countries that are having good socio-economic and health factors
    - Total 4 countries are in this category - belgium, ireland, mali and nepal
  - We have seen from both k-means and hierarchical clustering methods that hierarchical clustering selects an extra 99 countries. I would choose the final countries from hierarchical clustering as it gave accurate output than k-means clustering. I have compared the clusters and visualized from both methods and hierarchical clustering gave precise information than k-means clustering.

CONCLUSION

- Among the conclusions drawn from approach 1 (including ouliers) and approach 2 (excluding outliers), approach 1 seems to be the more appropriate choice because it includes all the data points including the outliers.
- As per the business requirements, we have to find all the countries which are in direst need of aid, i.E., Countries having low socio-economic and health factors. Hence we can't exclude any countries from our dataset as it will create a major drawback in our model.
- The final list of 48 countries name needs to focus on the most are mentioned below:
- Afghanistan, angola, benin, botswana, burundi, cambodia, canada, chad, chile, congo, dem. Rep., Congo, rep., Costa rica, croatia, equatorial guinea, eritrea, gabon, gambia, ghana, guinea, guinea-bissau, iceland, iraq, kenya, kiribati, lao, lesotho, liberia, madagascar, malawi, mali, mauritania, mozambique, namibia, niger, nigeria, pakistan, rwanda, senegal, sierra leone, solomon islands, south africa, sudan, tanzania, timor-leste, togo, uganda, yemen, zambia

THANK YOU