Selecting Countries in need of Aid for HELP International

Nitanshu Joshi

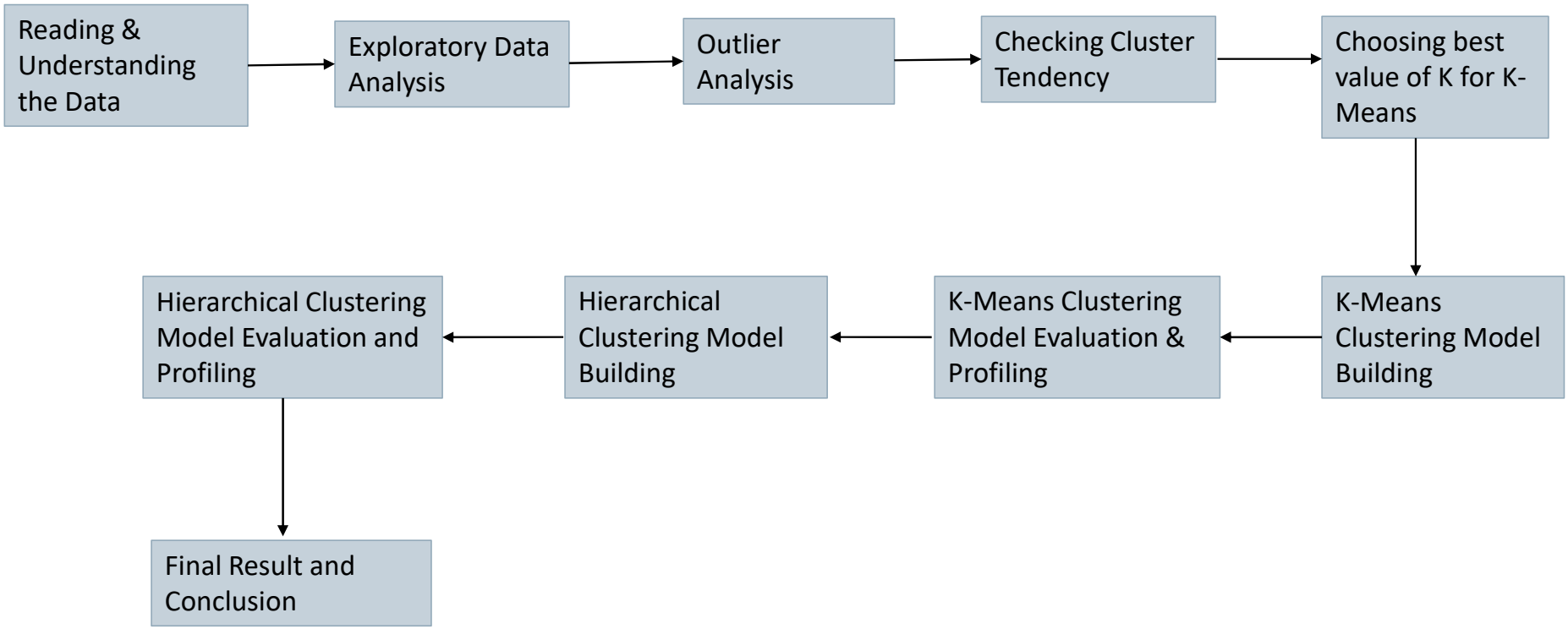
Problem Statement

HELP International is an international humanitarian NGO that is committed to fighting poverty and providing the people of backward countries with basic amenities and relief during the time of disasters and natural calamities.

They have been able to raise an amount of 10 Million Dollars for this purpose.

The objective of this assignment is to provide the CEO of HELP International a list of countries that are in direst need of this aid.

Analysis Approach



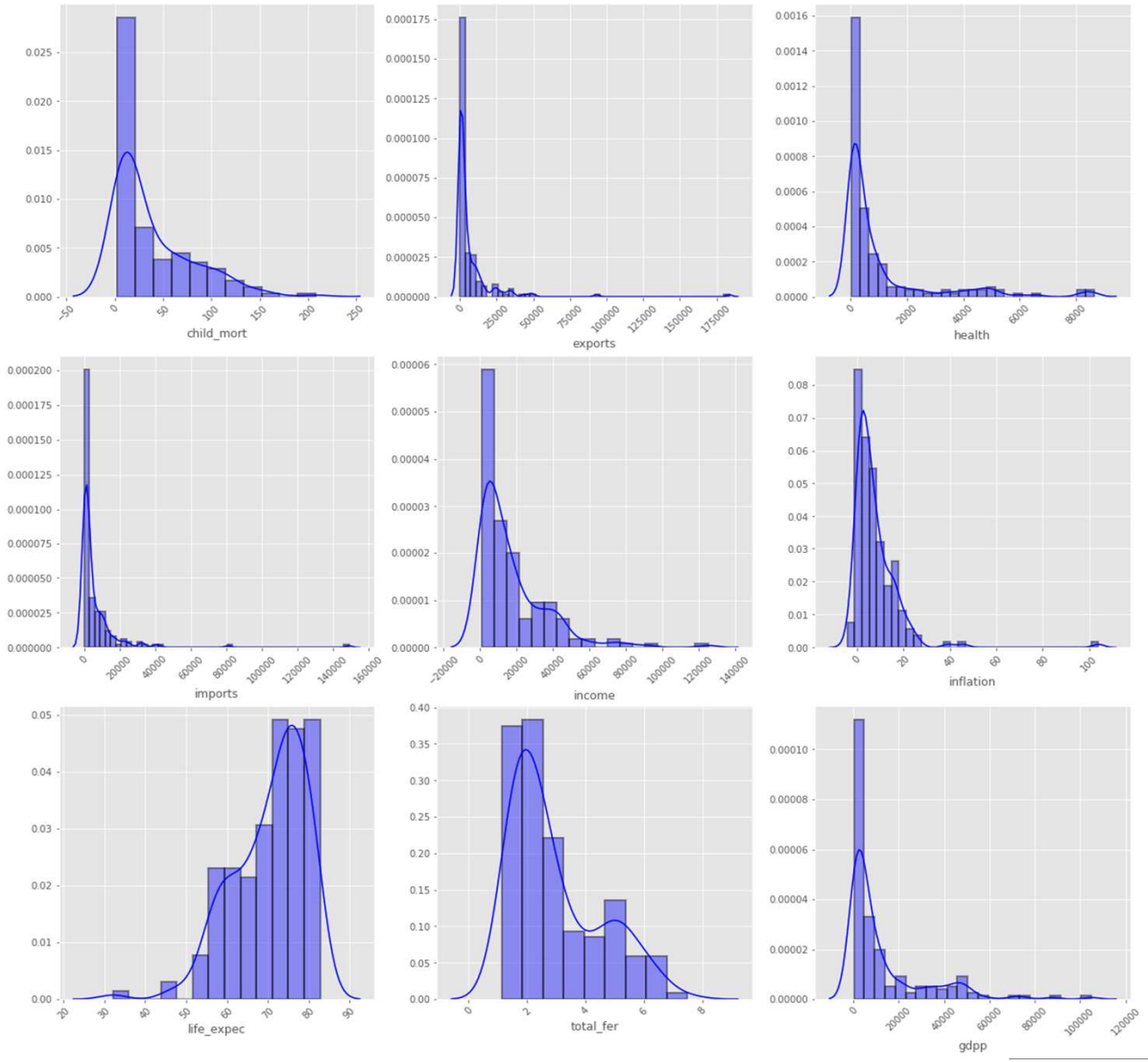
Data Dictionary

Column Name	Description
country	Name of the country
child_mort	Death of children under 5 years of age per 1000 live births
exports	Exports of goods and services per capita. Given as %age of the GDP per capita
health	Total health spending per capita. Given as %age of GDP per capita
imports	Imports of goods and services per capita. Given as %age of the GDP per capita
Income	Net income per person
Inflation	measurement of the annual growth rate of the GDP deflator
life_expec	The average number of years a new born child would live if the current mortality patterns are to remain the same
total_fer	The number of children that would be born to each woman if the current age-fertility rates remain the same.
gdpp	The GDP per capita. Calculated as the Total GDP divided by the total population.

Univariate Analysis of the Variables

From the Distplot following points can be concluded –

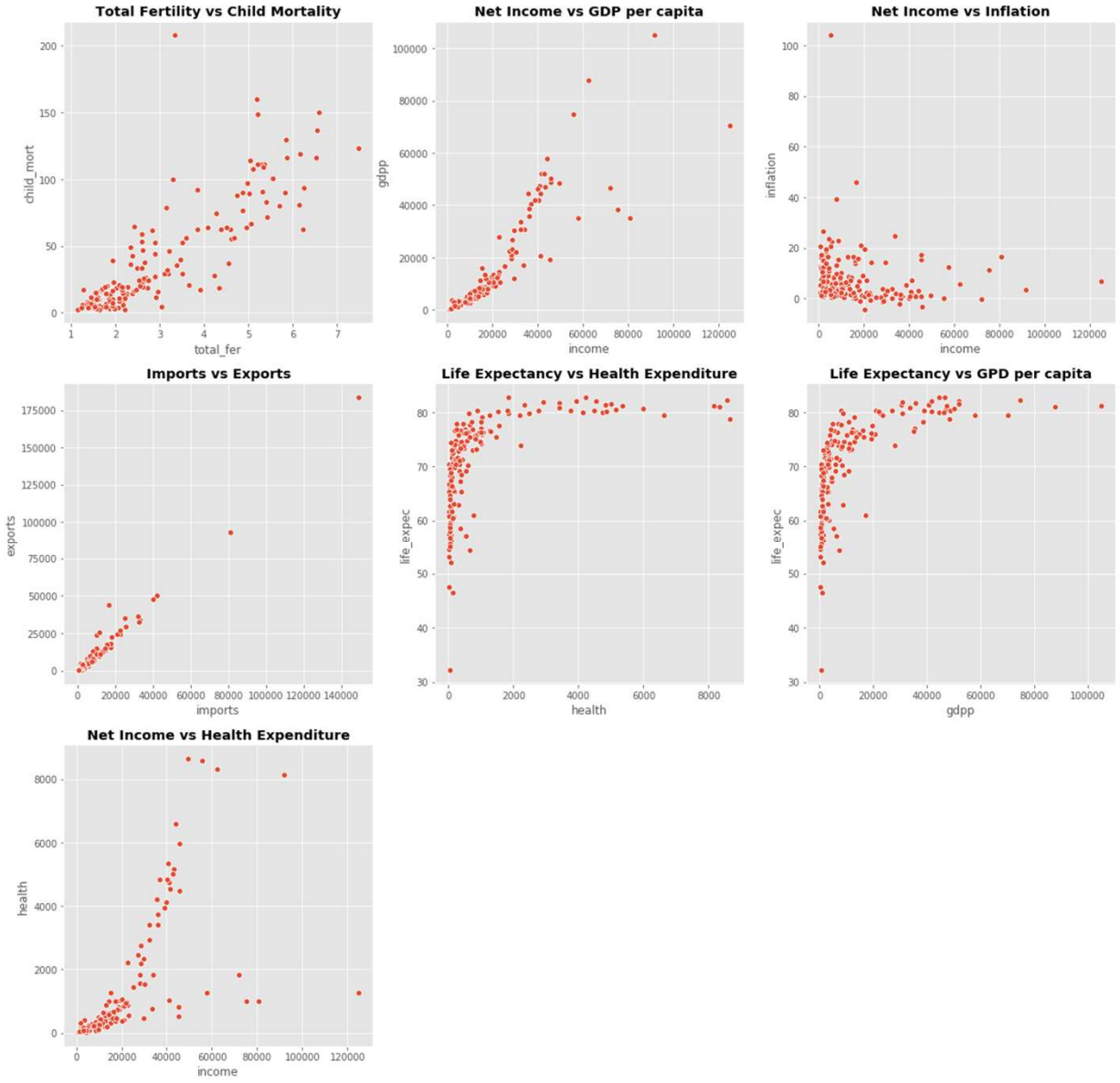
- All the variables except life_expec shows right-skewed behavior on the distplot. life_expec on the other hand shows left-skewed behavior.
- It can be inferred that life expectancy is high for most of the countries.
- Also, the other remaining columns - 'income', 'total_fer', 'gdpp', 'child_mort', 'inflation', 'imports', 'exports' and 'health' gives an inference that most of the countries have low values of the these.



Bivariate Analysis of the Variables

We observe the following from the various scatter plots:

- As the Total Fertility rates for a country increases, the child mortality rates also increases.
- A country with high GDPP will also have a higher Net income.
- We don't observe a linear relationship between income and inflation.
- We can say that higher the number of imports, higher will be the number of exports for a country.
- If the health expenditure increases, there is a high chance of life expectancy increasing too.
- The same can be said for gdpp vs life_expec plot. If the gdpp increases, there is a high chance of life expectancy increasing too.
- We can say that countries with high income tend to spend more on health expenditure.



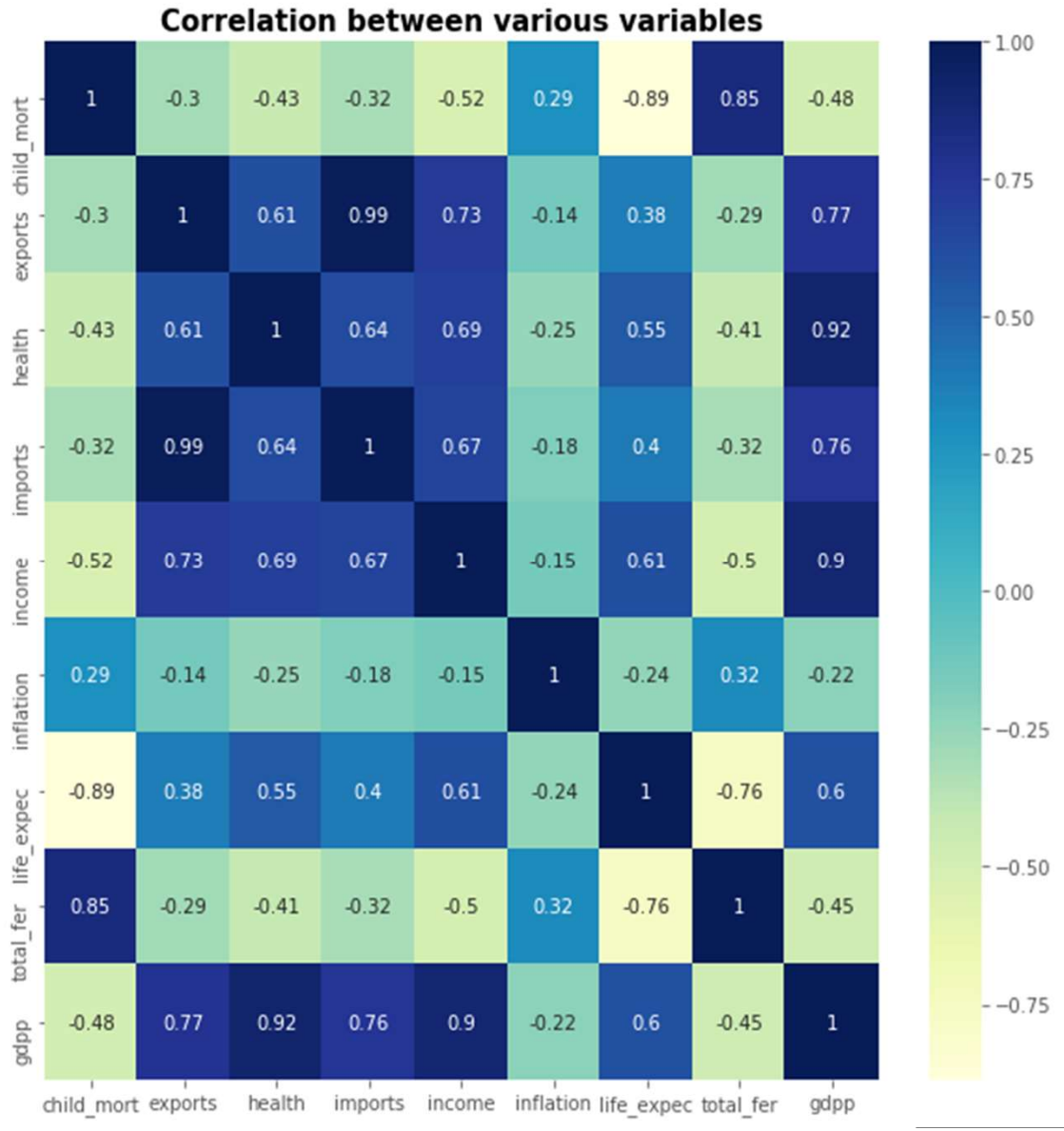
Visualizing Correlation in the Data

We observe that there is pretty high positive or negative correlation between most of the variables.

But, since Clustering is not affected much by collinearity,

Thus, we will ignore the collinearity between the variables for this case.

Hence no change or treatment of data is required here.



Outlier Analysis and Treatment

Observations:

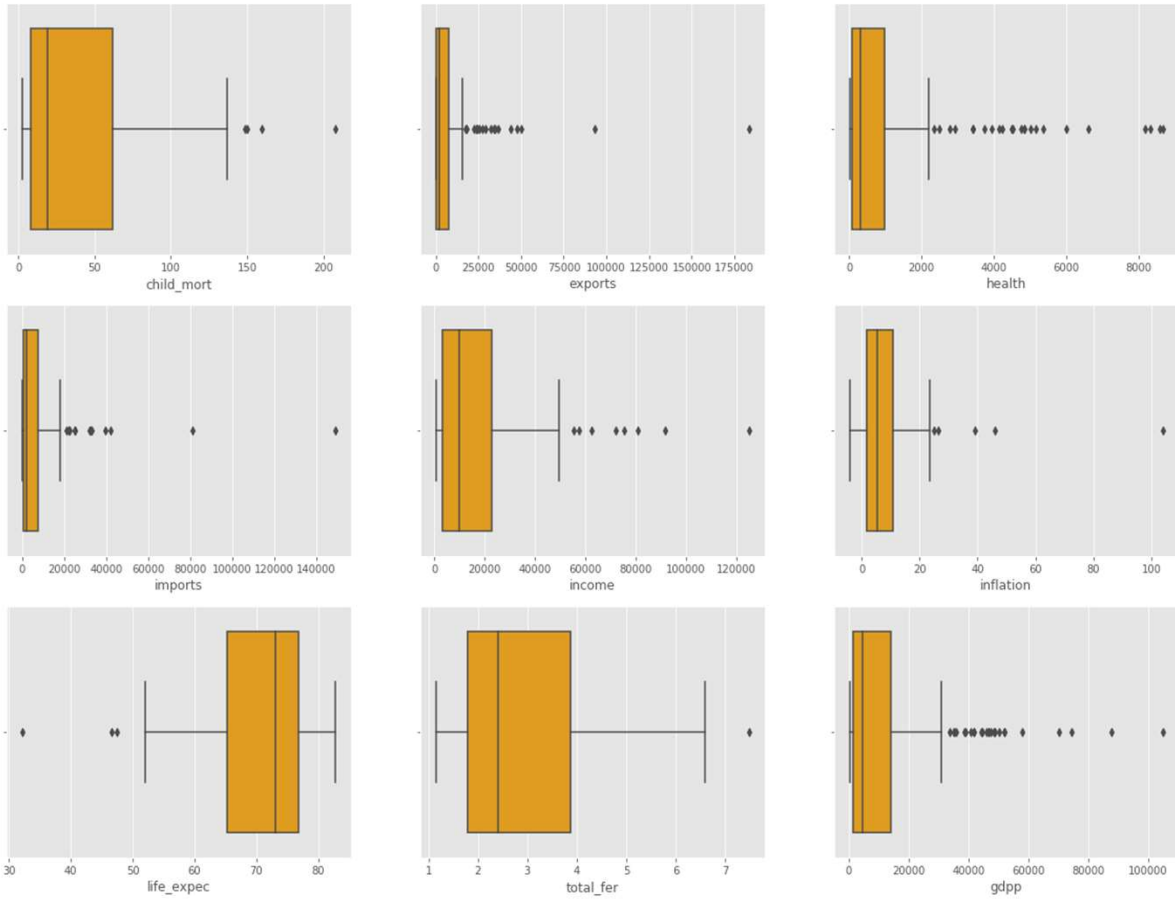
child_mortality, exports, imports,inflation , income, gdpp and health spending have some upper range outliers, whereas life_expentancy has a some lower range outliers

Outlier Analysis:

- For columns such as child_mort, inflation, total_fer we will not do anything to the upper range outliers but we will deal with the lower range outlier (capping).
- But for rest of the columns, we will not do anything for the lower range outliers but we will deal with the upper range outliers(capping).

Outlier Treatment:

- For child_mort, inflation, total_fer, life_expec ==> We will leave the outliers as it is.
- For gdpp, income, imports, exports, health ==> We will capp the above 99% of values



Checking Cluster Tendency (Hopkin's Test)

To check the tendency of a cluster we use the Hopkin's Statistics test. The higher the value on this test statistic, the better the data is for clustering.

If the result of Hopkin's statistics is more than 0.80, the data can be considered good for clustering. It is also a good practice to run the Hopkin's Test around 10 times to get a better output.

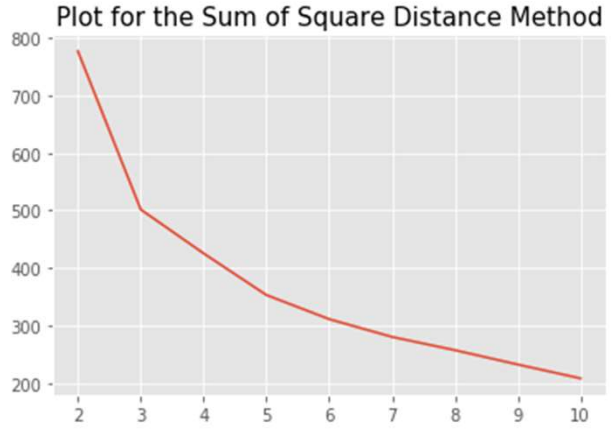
For our data, most of the values we got were more than 0.90 with an average of 0.91.

```
Running Hopkin's for the 0 time we get - 0.9165627606836735
Running Hopkin's for the 1 time we get - 0.9287413155508766
Running Hopkin's for the 2 time we get - 0.9341486471654658
Running Hopkin's for the 3 time we get - 0.9581987670126503
Running Hopkin's for the 4 time we get - 0.9307868896913247
Running Hopkin's for the 5 time we get - 0.9202295997811648
Running Hopkin's for the 6 time we get - 0.9405577402770081
Running Hopkin's for the 7 time we get - 0.9434832043520615
Running Hopkin's for the 8 time we get - 0.8722324261683153
Running Hopkin's for the 9 time we get - 0.92901657129846
```

Thus we can say that our data is suitable for clustering.

Choosing the Value of K

SSD (Elbow Curve) Method



We get a value of K = 3 by looking at this curve.

The Silhouette Score Method

	k	silhouettes_score
0	2	0.499264
1	3	0.424558
2	4	0.425873
3	5	0.391908
4	6	0.305271
5	7	0.273231
6	8	0.302369
7	9	0.298880
8	10	0.291787

By looking at the scores, we find the value of k = 4 to be the highest. Thus we get a value of k = 4. But there is not much difference in the scores for k=3 and k=4. Thus we can also consider k=3.

Final Selection of Value of K

From both the above methods we see that k=3 can be considered for our assignment.

K-Means Model with K=3

Income vs gdpp plot

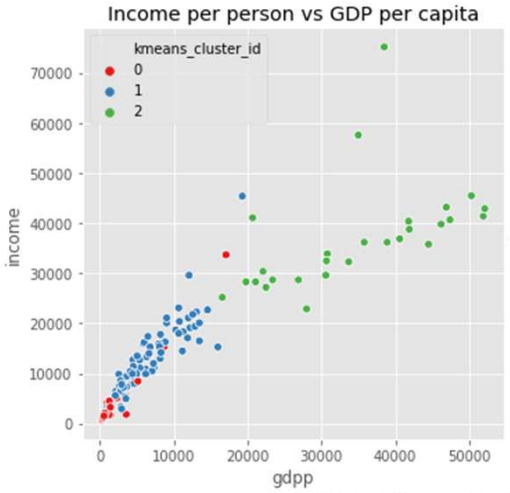
cluster 0 – have data points with generally low income and low gdpp
cluster 1 – have data points with generally medium income and medium gdpp.
cluster 2 – have data points with generally high income and high gdpp

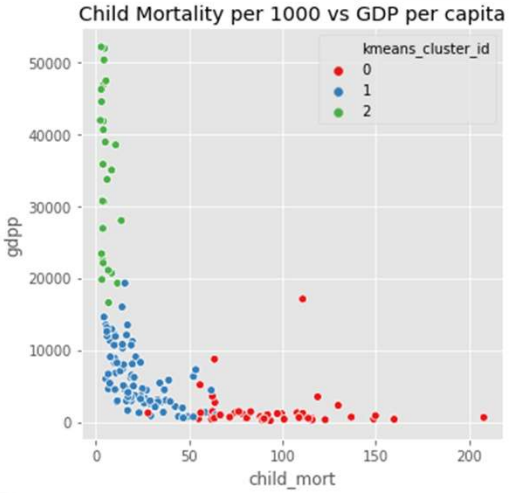
gdpp vs child_mort plot

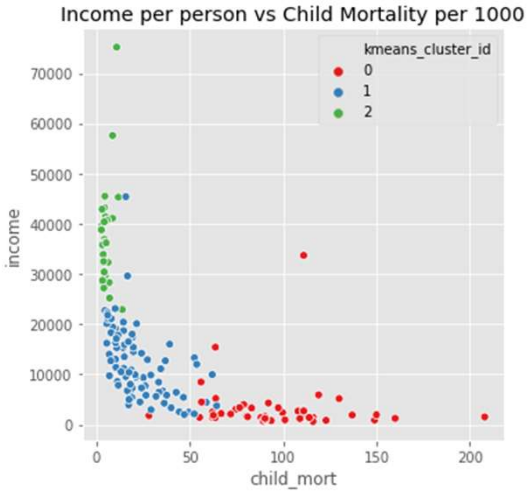
cluster 0 – data points with generally low gdpp and high child mortality rate.
cluster 1 – have data points with generally low-medium gdpp and medium child mortality rate.
cluster 2 – have data points with generally high gdpp and low child mortality rate.

income vs child_mort plot

cluster 0 – have data points with generally low income and high child mortality rate.
cluster 1 – have data points with generally low-medium income and medium child mortality rate.
cluster 2 – have data points with generally high income and low child mortality rate.







K-Means – Box Plot to find out the cluster of interest

From the shown boxplots we can conclude that –

Cluster 0 - Very high child mortality rate, very low GDP per capita, very low net income per individual.

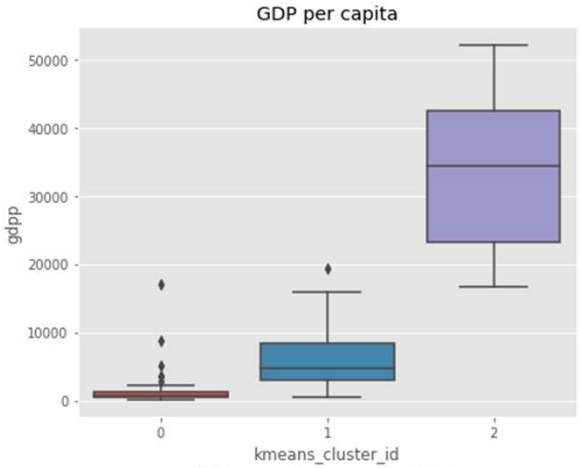
Cluster 1 - Medium child mortality rate, medium GDP per capita, medium net income per individual.

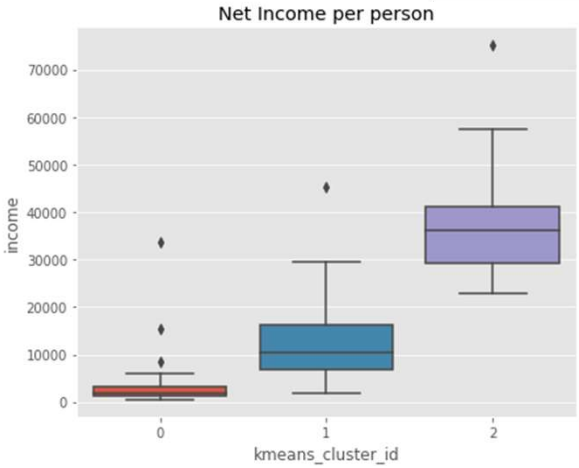
Cluster 2 - Very low child mortality rate, very high GDP per capita, very high net income per individual.

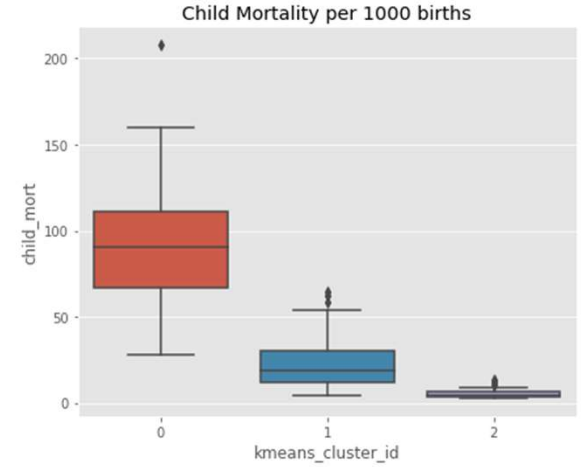
We have to select the cluster with countries having:

- High Child Mortality Rate per 1000 births.
- Low Net Income per individual.
- Low GDP per capita.

Thus, The cluster that we require is the Cluster Number = 0







Result from the K-Means Algorithm

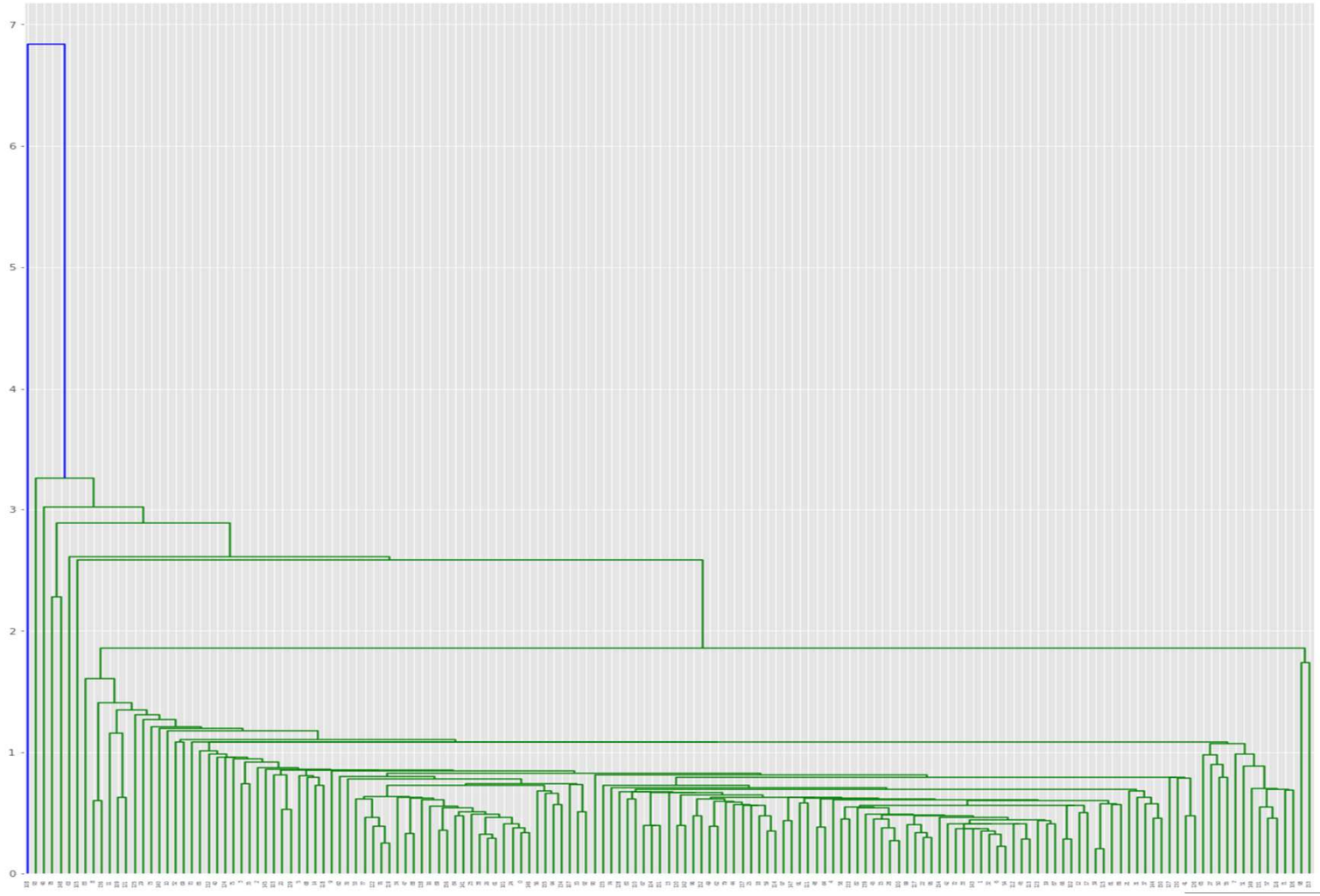
On analyzing the cluster number 0, we found the total number of countries present to be 45. When we sort the countries of cluster number 0 in ascending order of GDP per capita and Income and descending order of child mortality rate we get the countries in the following order.

The Top 20 countries in dire need of aid are –

1 - Burundi
2 - Liberia
3 - Congo, Dem. Rep.
4 - Niger
5 - Sierra Leone
6 - Madagascar
7 - Mozambique
8 - Central African Republic
9 - Malawi
10 - Eritrea
11 - Togo
12 - Guinea-Bissau
13 - Afghanistan
14 - Gambia
15 - Rwanda
16 - Burkina Faso
17 - Uganda
18 - Guinea
19 - Haiti
20 - Tanzania

Hierarchical Clustering with Single Linkage

Single Linkage Hierarchical Clustering Dendrogram



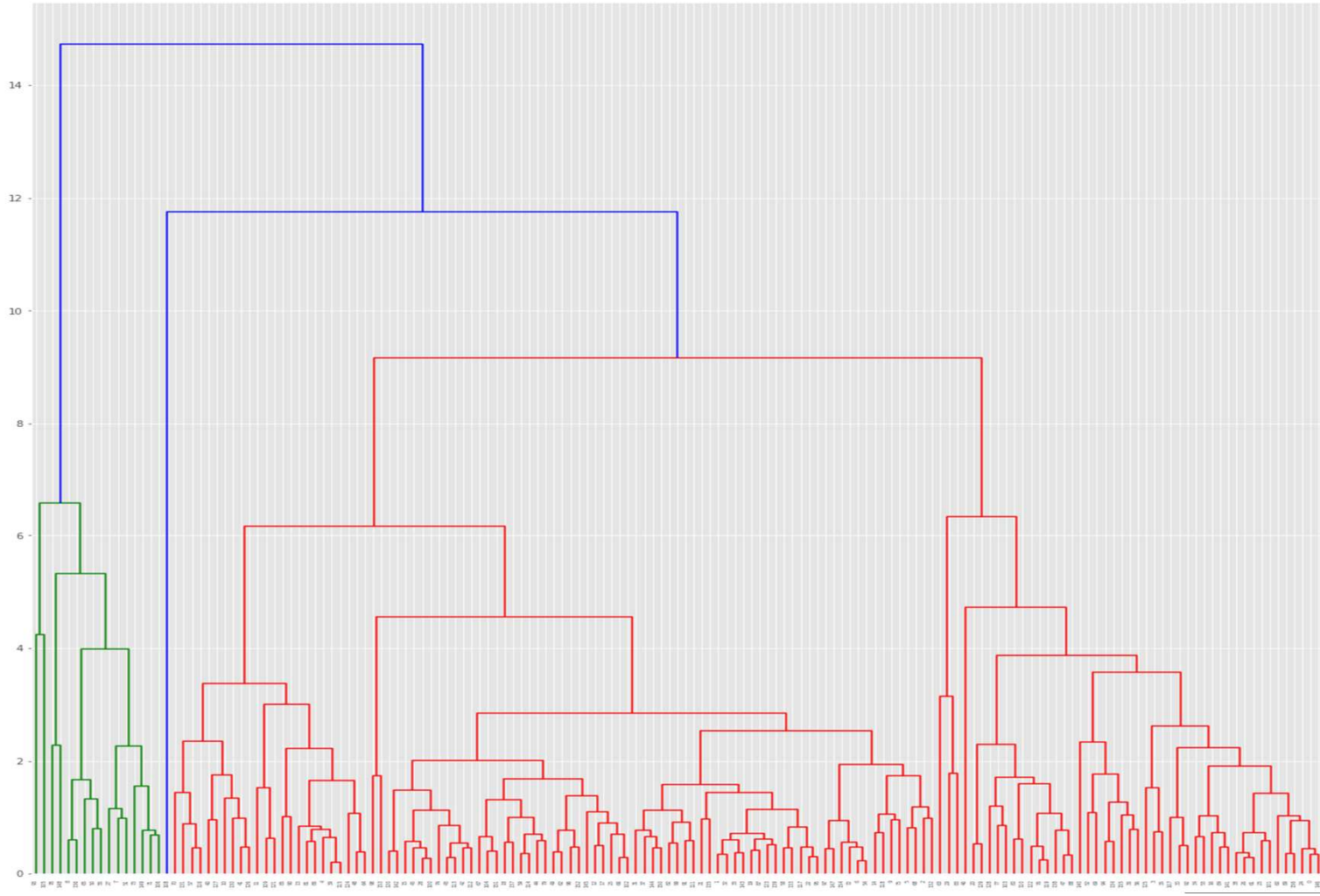
We observe that with Single Linkage method –

We get many clusters with only a single data point, making this method un-reliable and inefficient.

Thus we move on to Complete Linkage Hierarchical Clustering.

Hierarchical Clustering with Complete Linkage

Complete Linkage Hierarchical Clustering Dendrogram



Looking at the dendrograms for the complete linkage method we observe that -

The clusters here are much better distributed.

Thus we can move forward with clustering using the Complete Linkage Hierarchical Clustering Method.

Cutting the Dendrograms to obtain the Clusters

Here we will cut the clusters at both number of clusters = 3 and number of clusters = 4.

Then we will profile the clusters formed in both cases, and then select an optimal number of clusters.

We also used a little intuition for doing this, taking help from the dendrogram.

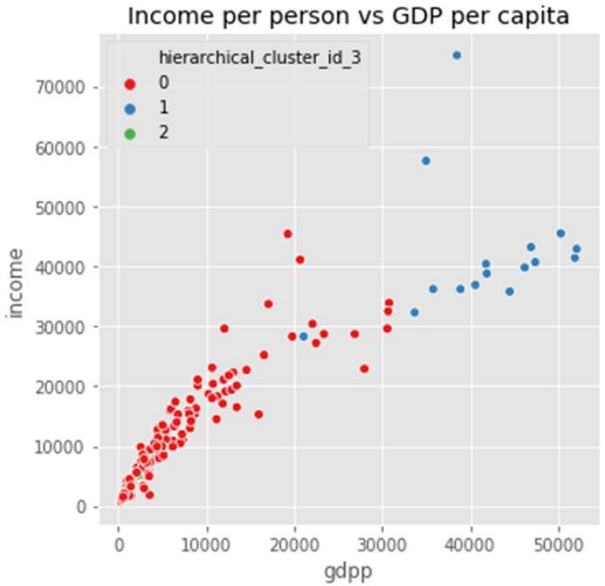
Cutting the Dendrograms at Number of Clusters = 3

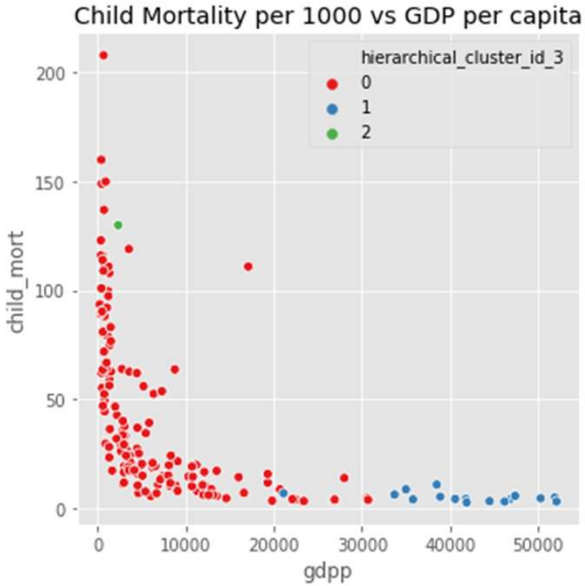
From the scatter plots we observe that –

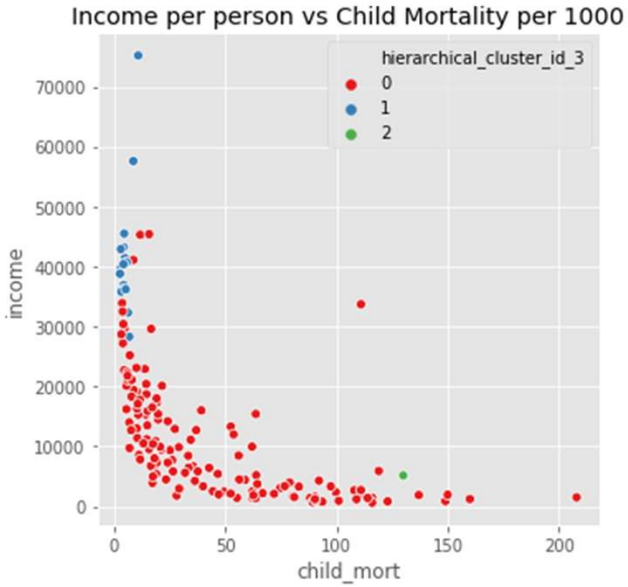
Cluster 0 have –
- Low to medium gddp and income
- Medium to high child mortality rate.

Cluster 1 have –
- Medium to high gddp and income.
- Have low child mortality rate

Cluster 2 have ***only one data point*** with –
- Low gddp and income.
- High child mortality rate







Cutting the Dendrograms at Number of Clusters = 3

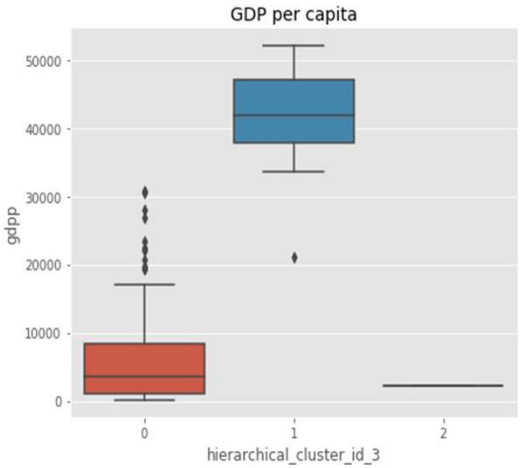
We will also look at the Boxplot for the various variables with respect to the clusters formed.

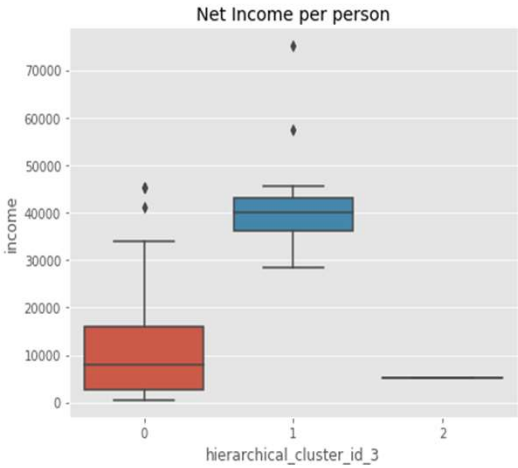
We observed that cluster 0 and cluster 2 are behaving similarly, while cluster 1 is entirely different.

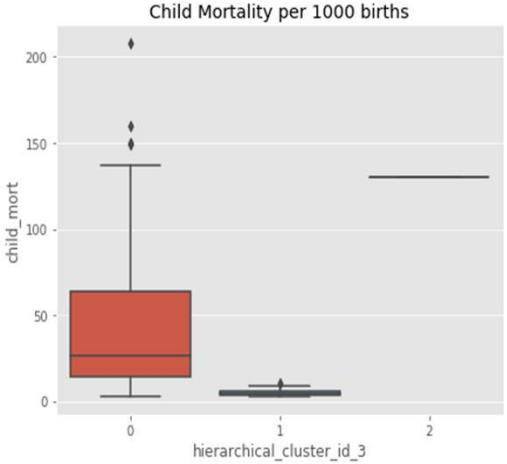
But on comparing cluster 0 with cluster 2 and 1, we see that -

- Cluster 0 - have a lower gdpp and lower income than cluster 2 & 1 when comparing the total data.
- Custer 0 have a higher child mortality rate than cluster 2 & 1 when comparing the total data.

The cluster that we require is the Cluster Number = 0







Issue with cutting the Dendrogram at Cluster size of 3

The issue with cutting the dendrogram at cluster size = 3 is that, the number of data points in this cluster which are 140 are extremely high. So it may not solve our Business needs.

Thus, to improve this we use cluster size of 4 to cut the dendrogram.

Cutting the Dendrograms at Number of Clusters = 4

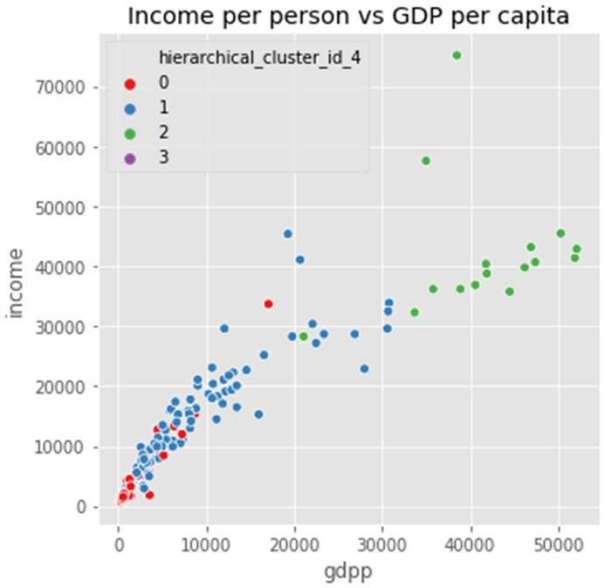
From the Cluster Plots we Observe that

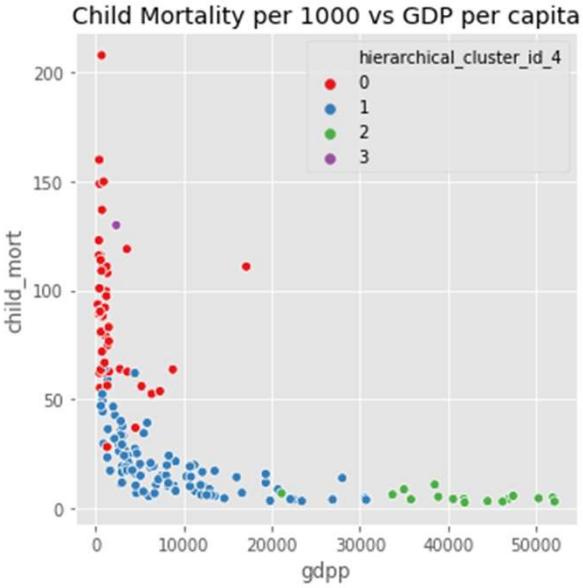
Cluster 0 have – Low Income, Low gddp, and Medium to High child mortality rate.

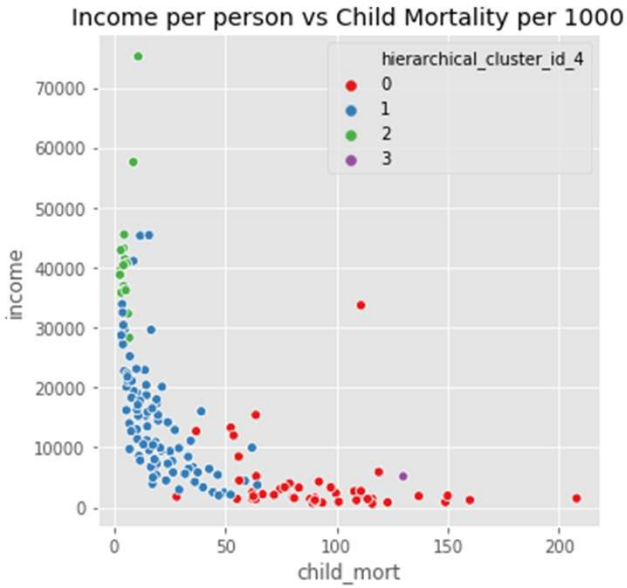
Cluster 1 have – Low to Medium Income and gddp, and Low child mortality rate.

Cluster 2 have – High income and gddp, and Low child mortality rate.

Cluster 3 has only 1 data point with low income and gddp, but with high child mortality rate.





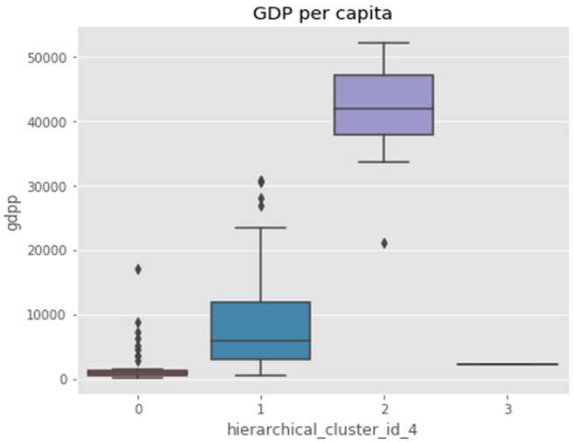


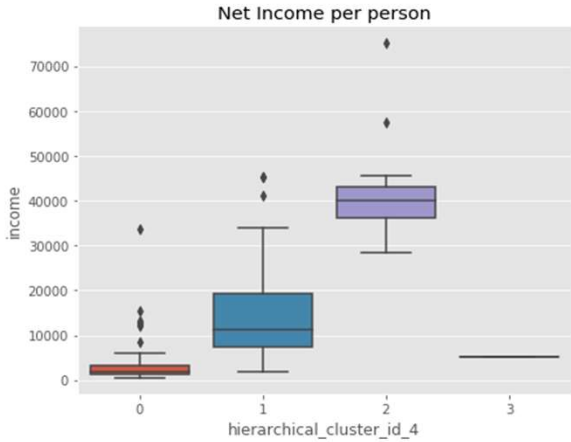
Cutting the Dendrograms at Number of Clusters = 4

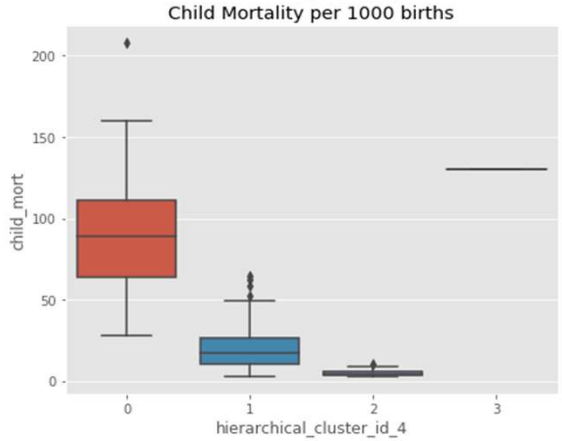
Cluster 0 - Low gdpp, Low income, High Child Mortality rate.
Cluster 1 - Medium gdpp, Medium income and low to medium child mortality rate.
Cluster 2 - High gdpp, High income, and very low child mortality rate.
Cluster 3 - Low gdpp, Low income, High Child Mortality rate

Here we see that cluster 0 and cluster 3 are nearly similar. but on comparing cluster 0 with cluster 3, we see that -
Cluster 0 have a lower gdpp & income than cluster 3 when comparing the total data. Also,
CLuster 0 have a higher child mortality rate than cluster 3 when comparing the total data.

The cluster that we require is the Cluster Number = 0.







Thus, we get a more optimal cluster distribution with cutting the dendrogram at 4 clusters.

Result from the Hierarchical Clustering Algorithm

On analyzing the cluster number 0, we found the total number of countries present to be 47.
When we sort the countries of cluster number 0 in ascending order of GDP per capita and Income and descending order of child mortality rate we get the countries in the following order.

The Top 20 countries in dire need of aid are –

1 - Burundi
2 - Liberia
3 - Congo, Dem. Rep.
4 - Niger
5 - Sierra Leone
6 - Madagascar
7 - Mozambique
8 - Central African Republic
9 - Malawi
10 - Eritrea
11 - Togo
12 - Guinea-Bissau
13 - Afghanistan
14 - Gambia
15 - Rwanda
16 - Burkina Faso
17 - Uganda
18 - Guinea
19 - Haiti
20 - Tanzania

Final Result and Conclusion

- We see that both the methods - K-Means Clustering and Hierarchical Clustering gives identical results.
- We found that K-Means generated a cluster of 45 countries and Hierarchical generated a cluster of 47 countries.
- So we can say that both the algorithms gave optimal results, although we had to take 3 clusters in K-means compared to 4 clusters in Hierarchical clustering.
- Both the choice of choosing number of clusters were verified using Silhouette's Score and not much difference was observed in choosing 3 clusters or 4 clusters.

The top 45 countries in need of humanitarian aid are -

1 - Burundi
2 - Liberia
3 - Congo, Dem. Rep.
4 - Niger
5 - Sierra Leone
6 - Madagascar
7 - Mozambique
8 - Central African Republic
9 - Malawi
10 - Eritrea
11 - Togo
12 - Guinea-Bissau
13 - Afghanistan
14 - Gambia
15 - Rwanda
16 - Burkina Faso
17 - Uganda
18 - Guinea
19 - Haiti
20 - Tanzania
21 - Mali
22 - Benin
23 - Comoros
24 - Chad
25 - Kenya
26 - Senegal
27 - Pakistan
28 - Lao
29 - Lesotho
30 - Mauritania
31 - Cote d'Ivoire
32 - Solomon Islands
33 - Cameroon
34 - Ghana
35 - Yemen
36 - Zambia
37 - Sudan
38 - Kiribati
39 - Congo, Rep.
40 - Angola
41 - Timor-Leste
42 - Iraq
43 - Namibia
44 - Botswana
45 - South Africa