Clustering Assignment

- GARGI SINGH

Objective

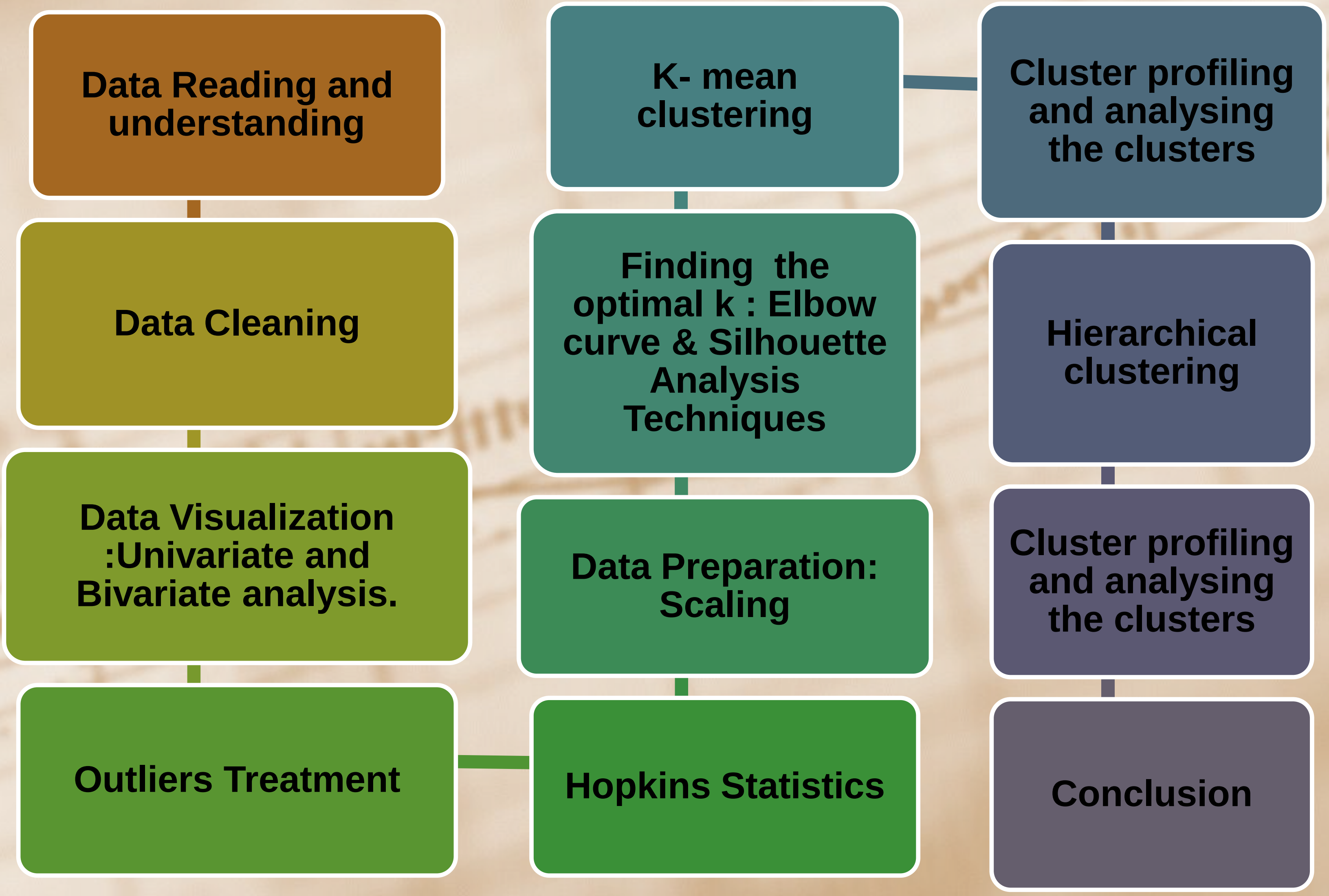
- To categorize the countries using some socio-economic and health factors that determine the overall development of the country.
- Then need to suggest the countries which the CEO needs to focus on the most.

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. It runs a lot of operational projects from time to time along with advocacy drives to raise awareness as well as for funding purposes.

After the recent funding programmes, they have been able to raise around $ 10 million. Now the CEO of the NGO needs to decide how to use this money strategically and effectively. The significant issues that come while making this decision are mostly related to choosing the countries that are in the direst need of aid.

Approach Applied :



Data Reading and understanding

Data Cleaning

Data Visualization :Univariate and Bivariate analysis.

Outliers Treatment

Hopkins Statistics

Data Preparation: Scaling

Finding the optimal k : Elbow curve & Silhouette Analysis Techniques

K- mean clustering

Cluster profiling and analysing the clusters

Hierarchical clustering

Cluster profiling and analysing the clusters

Conclusion

Data reading

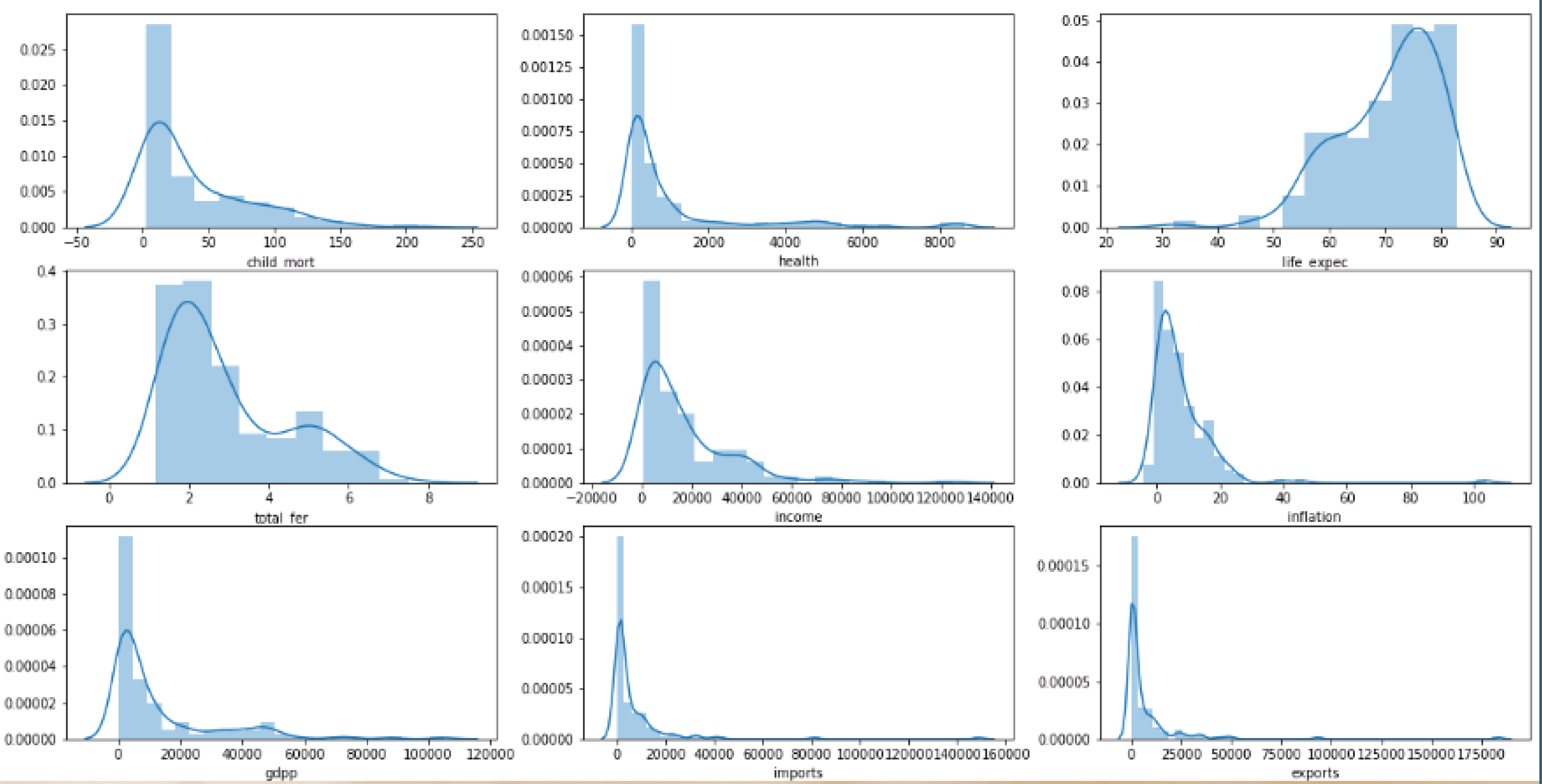
- Imported the data and viewed the dataset.
- Understood the type of data .
- Read the shape of the dataset.
- Checked few columns like export,import and health was in the percentage form changed to the actual value to give us the proper analysis of the data later.

Data Cleaning

- The dataset was clean enough.
- Checked the null values if any but found to be zero so decided to proceed further for the analysis part .
- Before getting into the next step also checked if any duplicate rows presents and concluded the dataset was free from the duplicate rows too.

Data Visualisation

Univariate Analysis



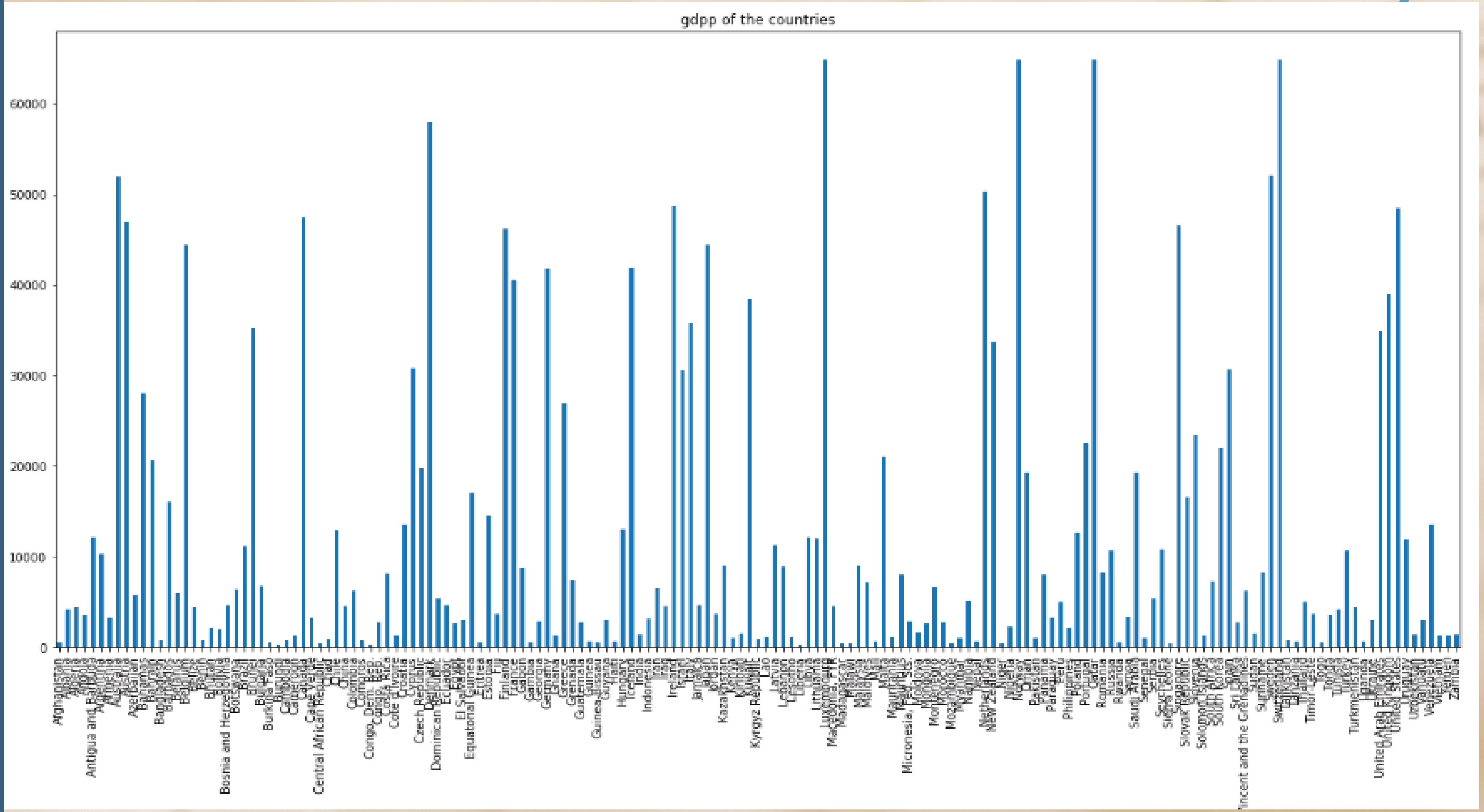
Data Visualization

Univariate Analysis

- There is some variation in the distribution of the income dataset .
- The child mortality also shows variation in the distribution of the dataset .
- The gdpp also shows the data distributed with some variation .
- We will takes these features for the data profiling for helping us know the countries in need of aid.

Data Visualization

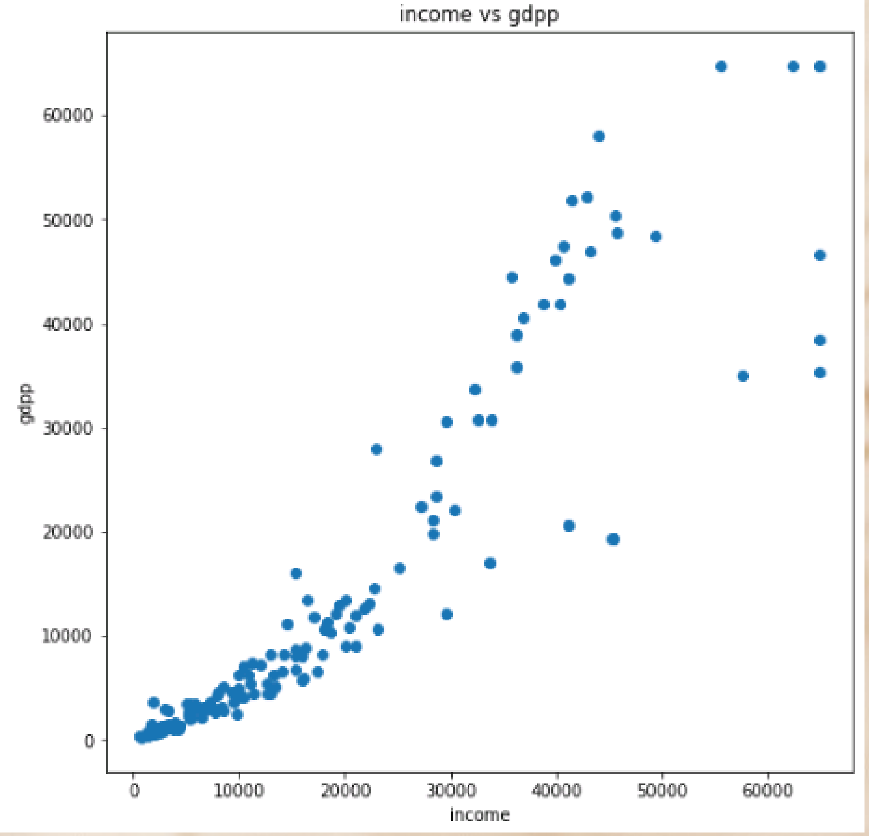
Univariate Analysis



We can see from the above gdpp bar plot there are good number of countries with high and low gdpp making good clusters and we can get countries for our purpose .

Data Visualization

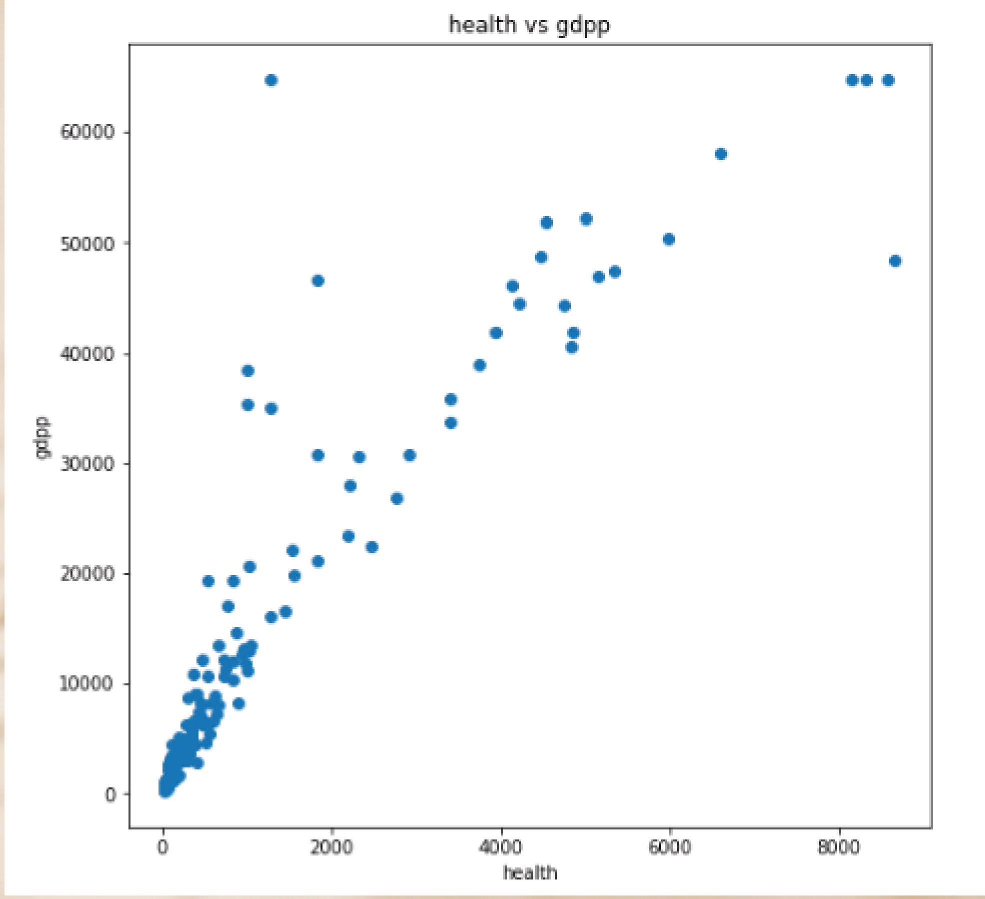
Bivariate Analysis



- We can notice from the plot that countries with the low income has low gdpp.
- We can focus on these countries.
- The countries with the low gdpp will definitely be in need of the aid .
- Countries with the high income rate shows good gdpp thus we can include them in the rich countries as of now .

Data Visualization

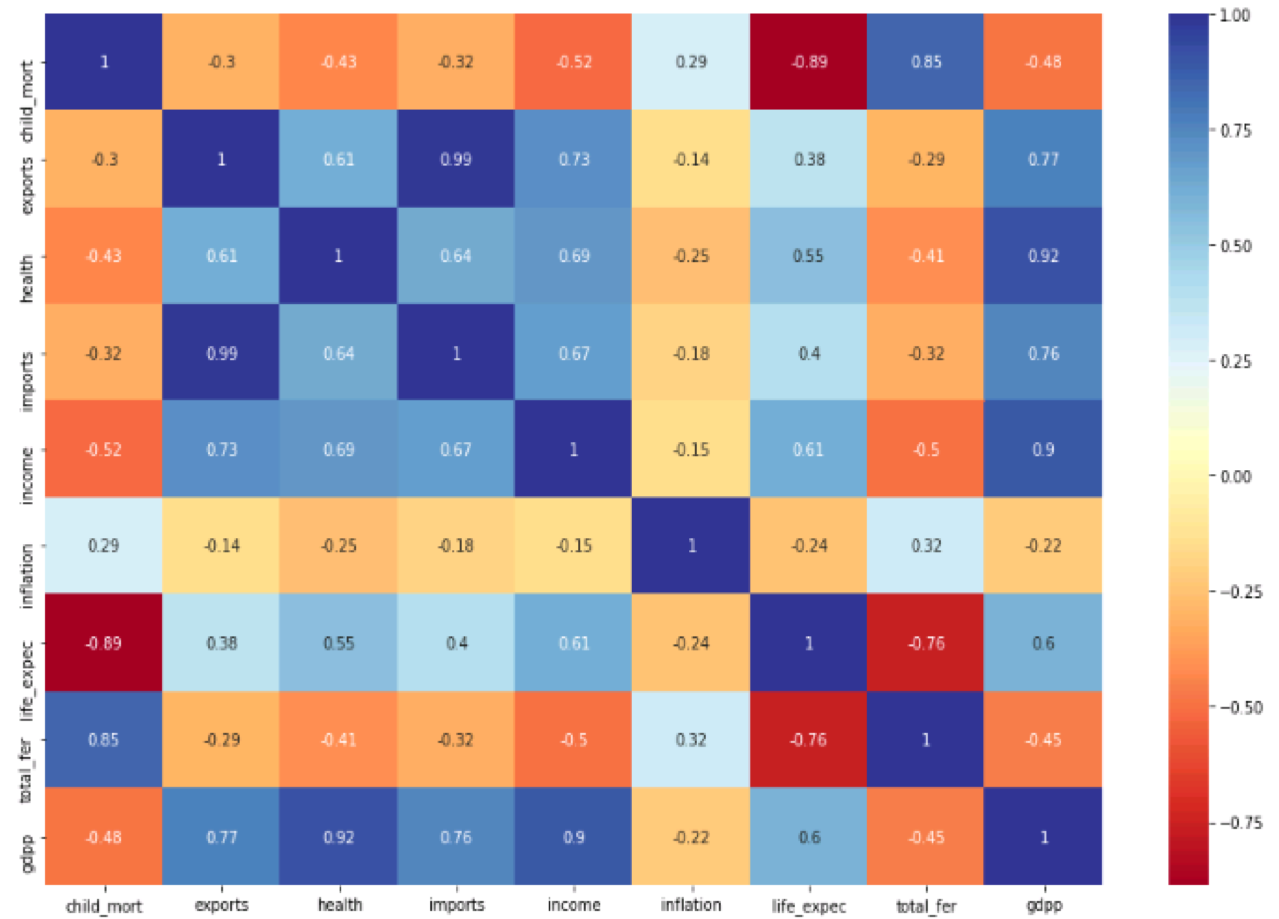
Bivariate Analysis



- We can notice from the plot that countries with the low gdpp are poor in the health also. We can focus on these countries.
- The countries with the low gdpp will definitely be in need of the aid .
- Countries with the high gdpp rate shows good health thus we can include them in the rich countries as of now .

Data Visualization

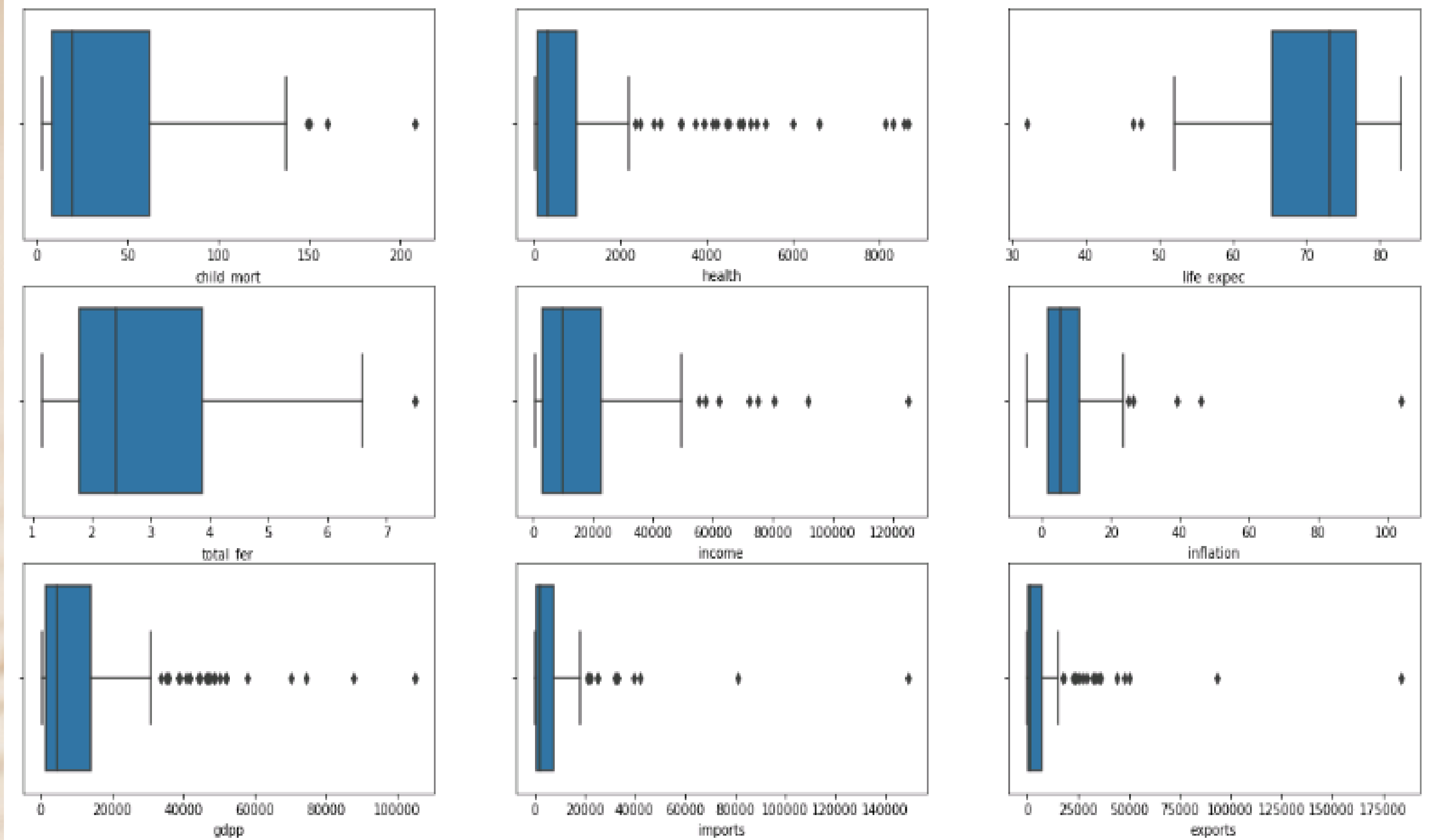
Correlation in the Data



	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
child_mort	1	-0.3	-0.43	-0.32	-0.52	0.29	-0.89	0.85	-0.48
exports	-0.3	1	0.61	0.99	0.73	-0.14	0.38	-0.29	0.77
health	-0.43	0.61	1	0.64	0.69	-0.25	0.55	-0.41	0.92
imports	-0.32	0.99	0.64	1	0.67	-0.18	0.4	-0.32	0.76
income	-0.52	0.73	0.69	0.67	1	-0.15	0.61	-0.5	0.9
inflation	0.29	-0.14	-0.25	-0.18	-0.15	1	-0.24	0.32	-0.22
life_expec	-0.89	0.38	0.55	0.4	0.61	-0.24	1	-0.76	0.6
total_fer	0.85	-0.29	-0.41	-0.32	-0.5	0.32	-0.76	1	-0.45
gdpp	-0.48	0.77	0.92	0.76	0.9	-0.22	0.6	-0.45	1

- The gdpp and the income had the high correlation
- Even the child_mort and the total_fer also shows high correlation
- The life_exp and the gdpp also shows a good correlation between each other
- The child_mort and the life_fer shows high negative correlation with each other

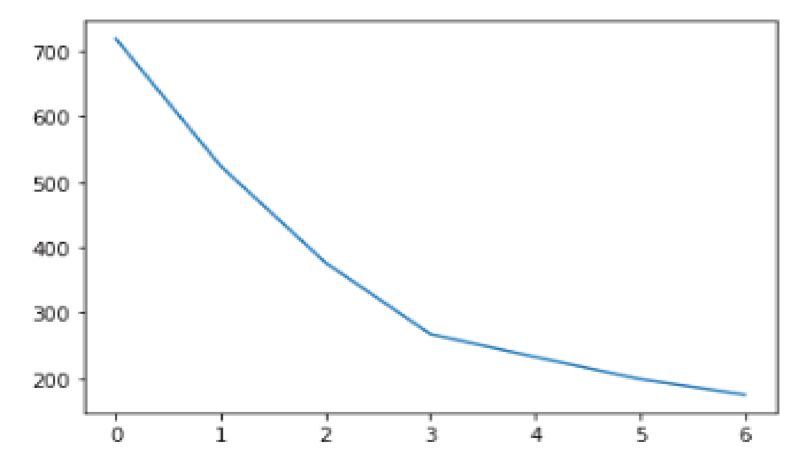
Outlier Treatment



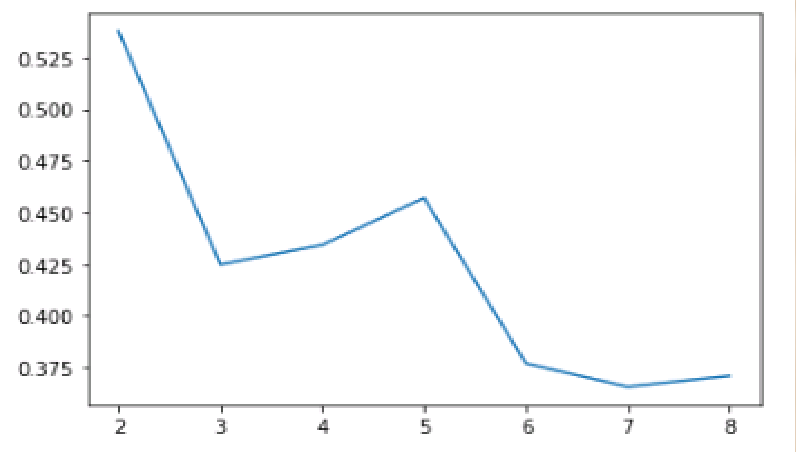
Treated few outliers which might cause change in our analysis by capping it .

K-mean clustering

Elbow Curve
(Sum of Squared distances)

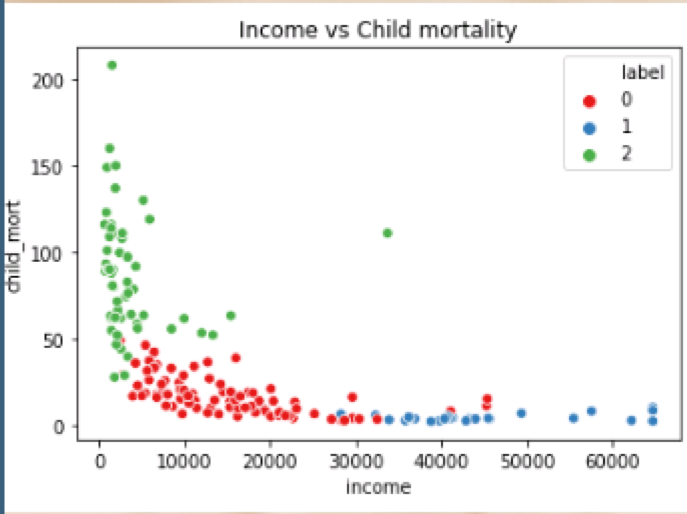


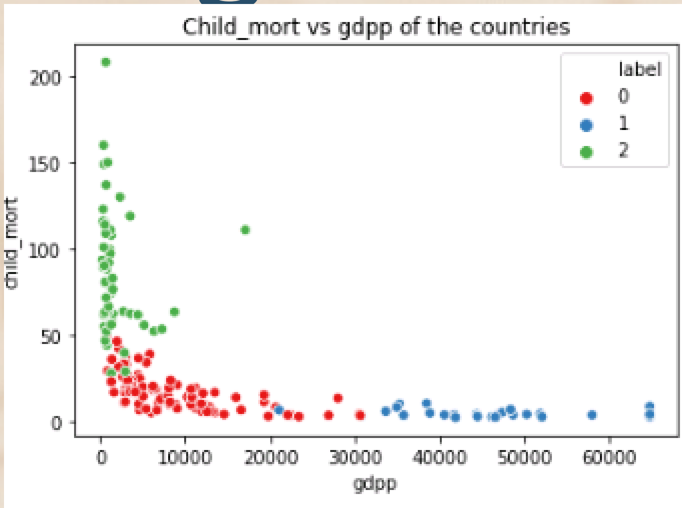
Silhouette Analysis

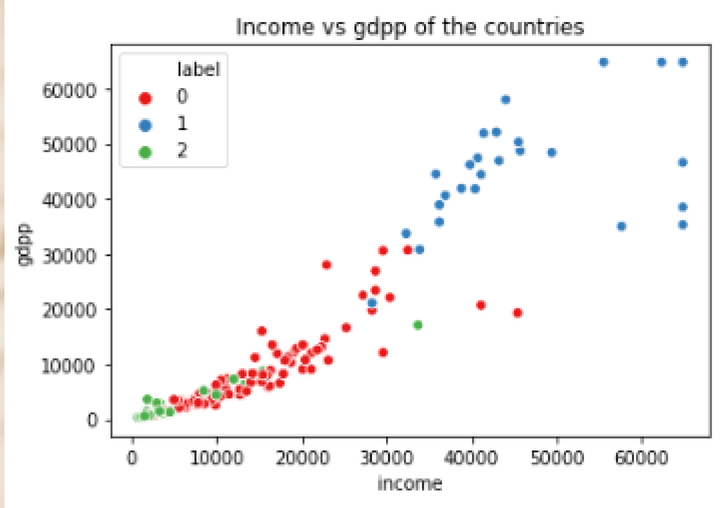


Looking at the above elbow curve and the Silhouette Analysis it is good to proceed with 3 clusters. As 4 clusters might confuse with the behaviors .

K-mean clustering


Income vs Child mortality


Child_mort vs gdpp of the countries


Income vs gdpp of the countries

Decided to go for 3 clusters where few things we observed:

- **The countries in green shows some behaviors like high child mortality , Low gdpp and low income .**
- **So the countries belonging to such clusters can be considered for the need of the aid.**

K-mean clustering

Fig:1

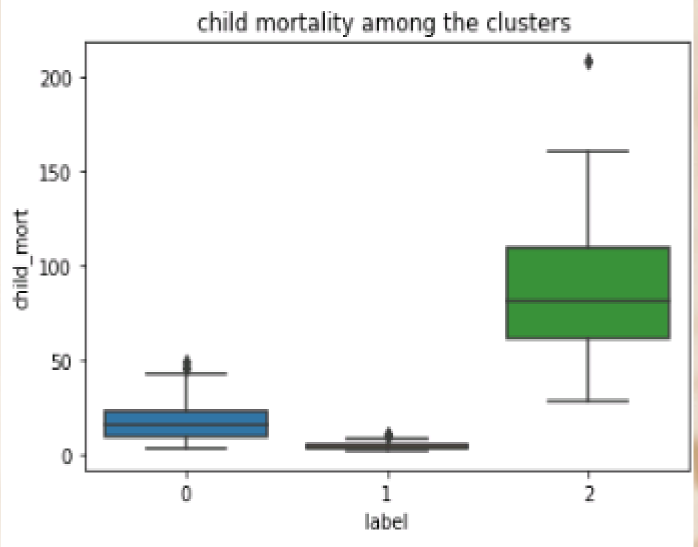
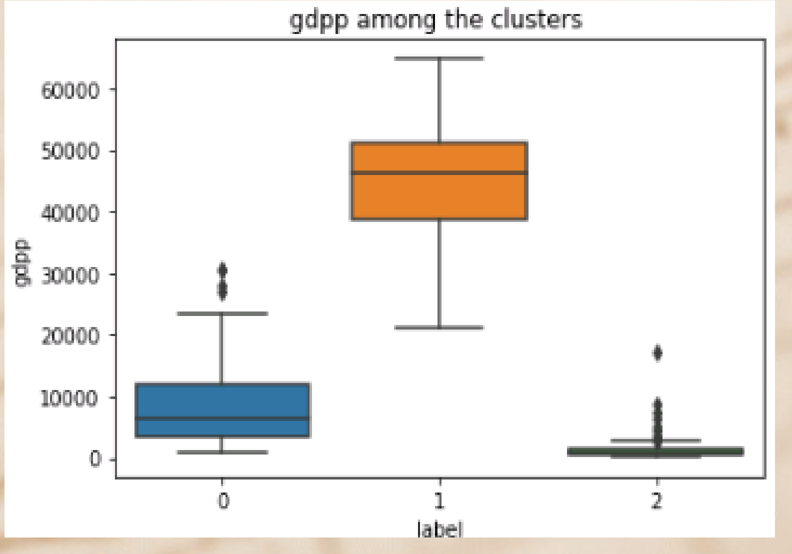


Fig:2



Things observed:

- **Fig1: The countries in cluster2 shows high child mortality . Child mortality indicates the Death of children under 5 years of age per 1000 live births so the cluster 2 is in more need of the aid.**

- **Fig2: The countries in the cluster 2 show lowest gdpp. The development of the countries of low gdpp are slower than the countries with the higher gdpp .Hence they are in need of the aid.**

K-mean clustering

Fig:3

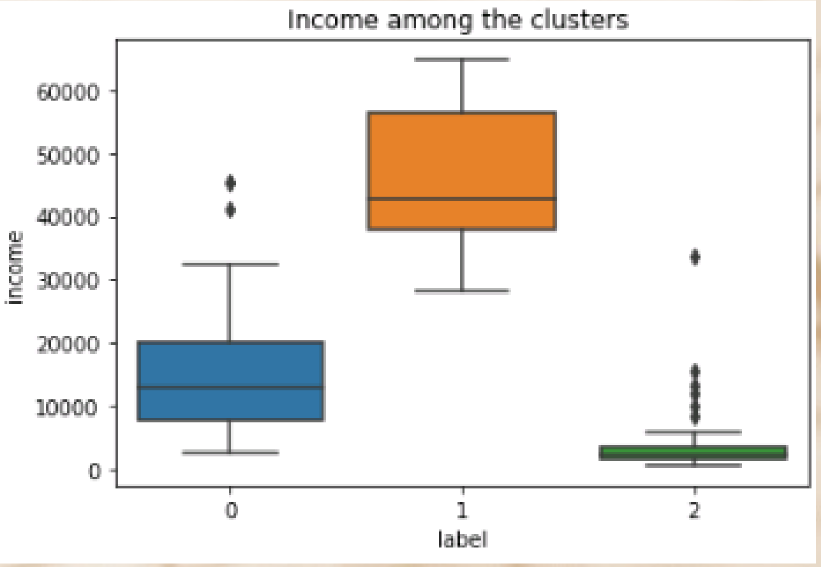
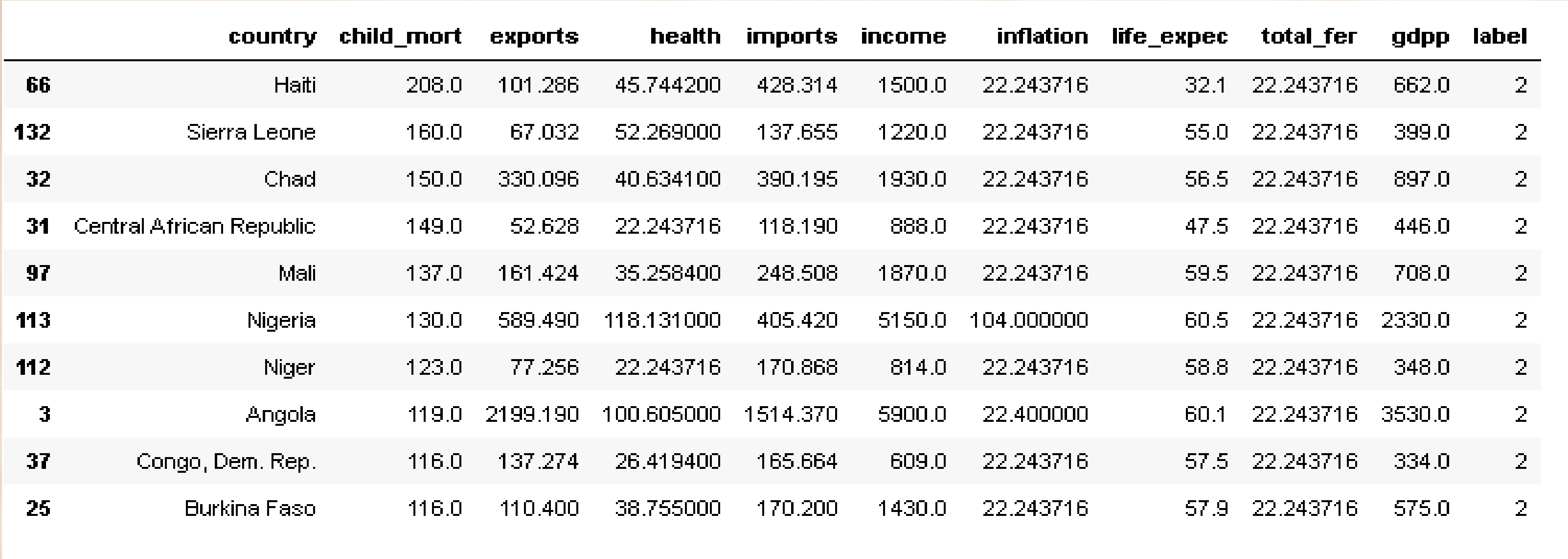


Fig3:

- **The countries belonging to the cluster 2 shows the behavior of low income.**
- **Hence countries coming under this cluster can be assumed as the poor countries and in need of the aid .**

K-mean clustering

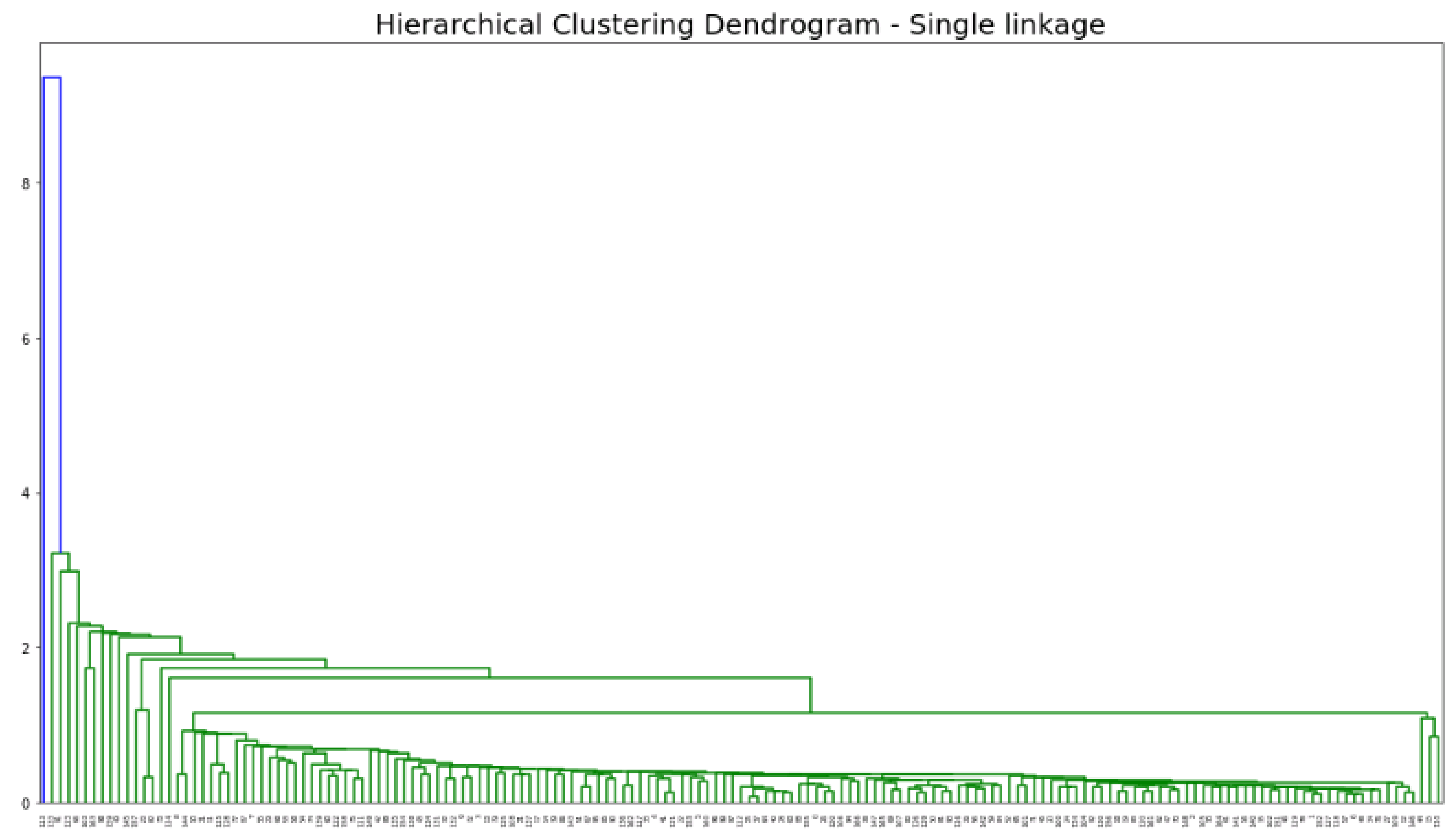


	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	label
66	Haiti	208.0	101.286	45.744200	428.314	1500.0	22.243716	32.1	22.243716	662.0	2
132	Sierra Leone	160.0	67.032	52.269000	137.655	1220.0	22.243716	55.0	22.243716	399.0	2
32	Chad	150.0	330.096	40.634100	390.195	1930.0	22.243716	56.5	22.243716	897.0	2
31	Central African Republic	149.0	52.628	22.243716	118.190	888.0	22.243716	47.5	22.243716	446.0	2
97	Mali	137.0	161.424	35.258400	248.508	1870.0	22.243716	59.5	22.243716	708.0	2
113	Nigeria	130.0	589.490	118.131000	405.420	5150.0	104.000000	60.5	22.243716	2330.0	2
112	Niger	123.0	77.256	22.243716	170.868	814.0	22.243716	58.8	22.243716	348.0	2
3	Angola	119.0	2199.190	100.605000	1514.370	5900.0	22.400000	60.1	22.243716	3530.0	2
37	Congo, Dem. Rep.	116.0	137.274	26.419400	165.664	609.0	22.243716	57.5	22.243716	334.0	2
25	Burkina Faso	116.0	110.400	38.755000	170.200	1430.0	22.243716	57.9	22.243716	575.0	2

Countries those are in the urgent need of the aid according to k-mean clustering:

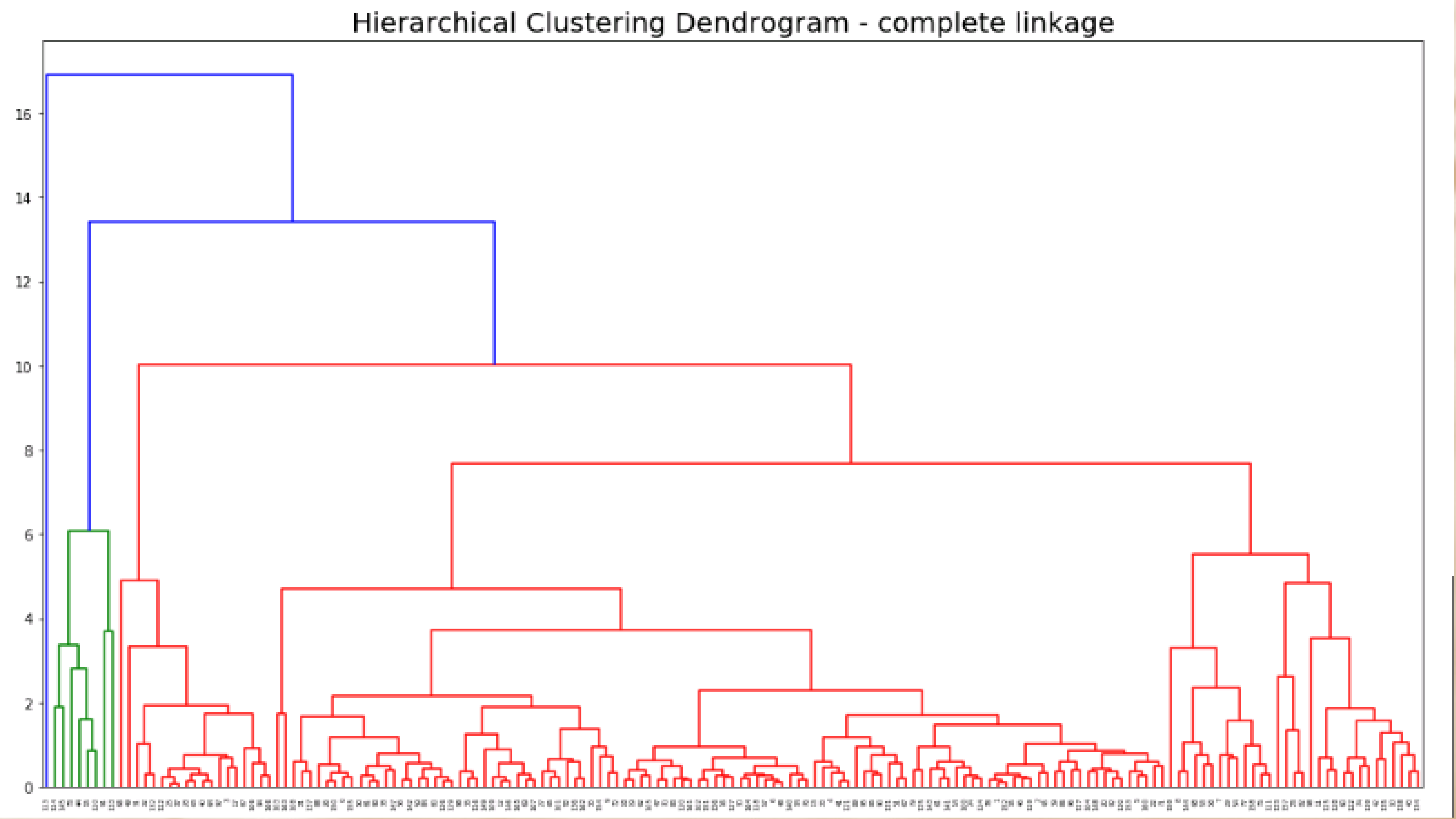
1. **Haiti**
2. **Sierra Leone**
3. **Chad**
4. **Central African Republic**
5. **Mali**
6. **Nigeria**
7. **Niger**
8. **Angola**
9. **Congo, Dem. Rep.**
10. **Burkina Faso**

Hierarchical clustering



Single linkage Hierarchical clustering was not very clear so we will look into the complete linkage and decide on the cluster for our analysis..

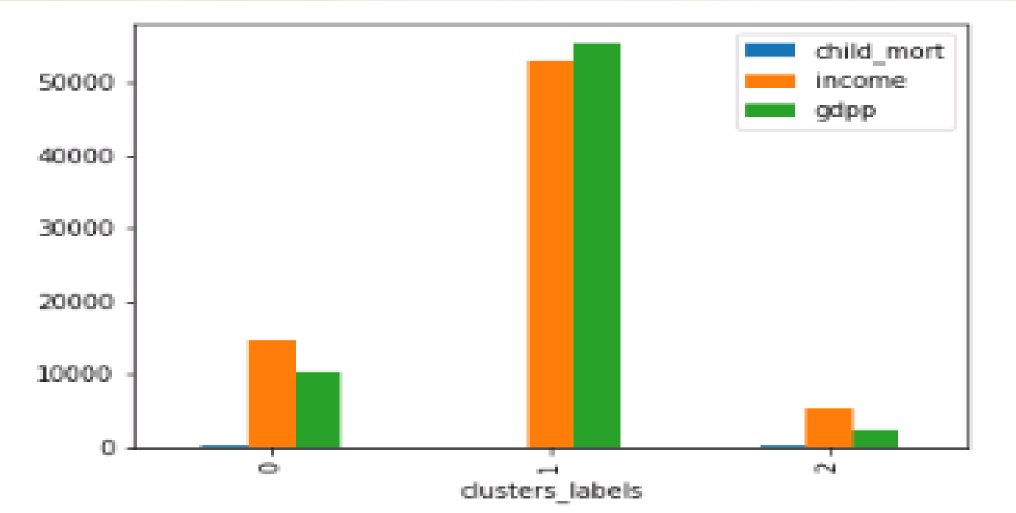
Hierarchical clustering



Complete linkage Hierarchical clustering gives us a option to refine the countries into many clusters and analyze .

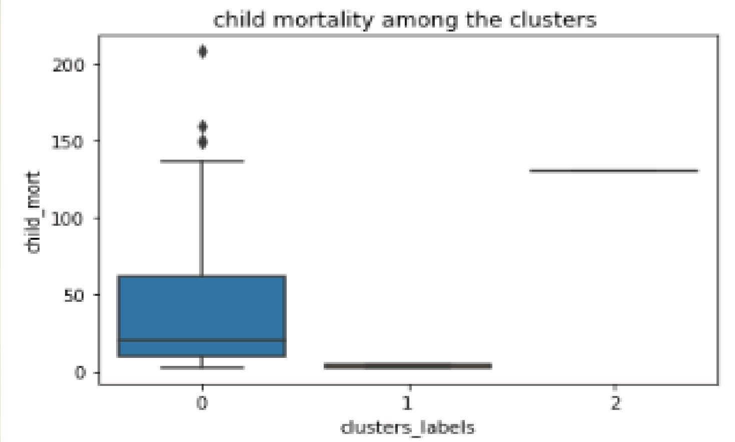
But making more clusters are going to confuse with the behavior of the countries so will go with 3 clusters only .

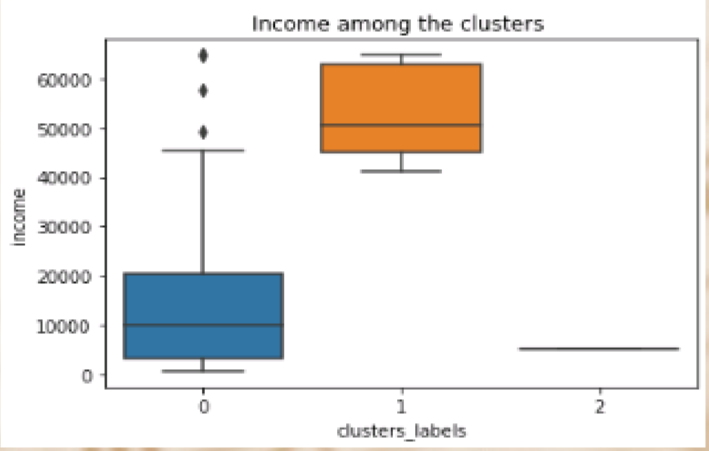
Hierarchical clustering

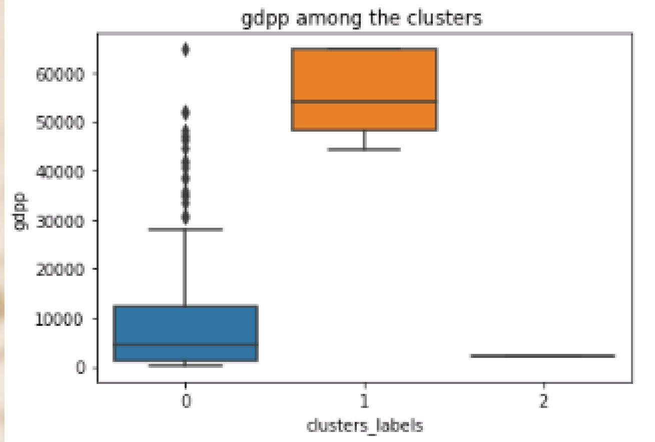


- The above are the countries for the features child mortality, gdpp and income.
- Child mortality is much smaller in scale compared to income and gdpp hence to verify the profile of child mortality against the cluster we will plot it separately and check.
- If we see we can make out from the above plot that the countries in shows the need of the aid at the most with the lower gdpp and income .

Hierarchical clustering



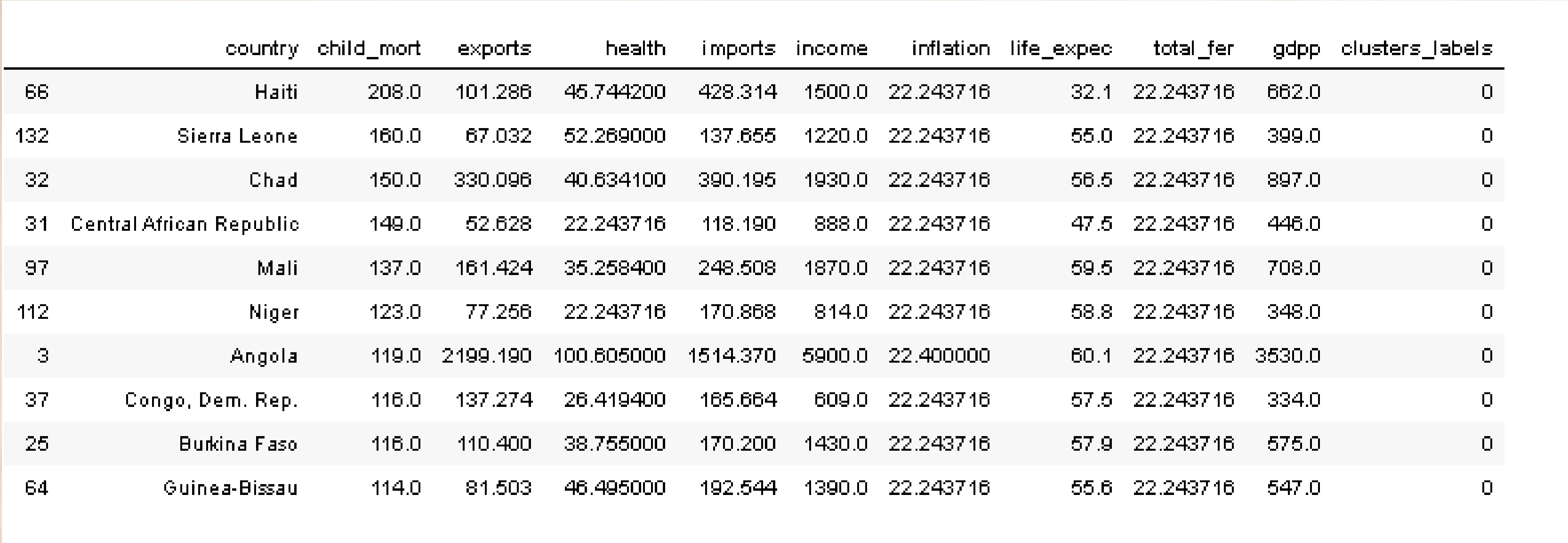




The countries in the cluster 0 shows the behavior of high child mortality, low income and low gdpp .So we can go ahead with the countries in this cluster .

There is only one country in the cluster 2 Nigeria which is definitely one of the country exhibiting the behavior of low income ,low gdpp and high child mortality.

Hierarchical clustering



	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	clusters_labels
66	Haiti	208.0	101.286	45.744200	428.314	1500.0	22.243716	32.1	22.243716	662.0	0
132	Sierra Leone	160.0	67.032	52.269000	137.655	1220.0	22.243716	55.0	22.243716	399.0	0
32	Chad	150.0	330.096	40.634100	390.195	1930.0	22.243716	56.5	22.243716	897.0	0
31	Central African Republic	149.0	52.628	22.243716	118.190	888.0	22.243716	47.5	22.243716	446.0	0
97	Mali	137.0	161.424	35.258400	248.508	1870.0	22.243716	59.5	22.243716	708.0	0
112	Niger	123.0	77.256	22.243716	170.868	814.0	22.243716	58.8	22.243716	348.0	0
3	Angola	119.0	2199.190	100.605000	1514.370	5900.0	22.400000	60.1	22.243716	3530.0	0
37	Congo, Dem. Rep.	116.0	137.274	26.419400	165.664	609.0	22.243716	57.5	22.243716	334.0	0
25	Burkina Faso	116.0	110.400	38.755000	170.200	1430.0	22.243716	57.9	22.243716	575.0	0
64	Guinea-Bissau	114.0	81.503	46.495000	192.544	1390.0	22.243716	55.6	22.243716	547.0	0

1. **Haiti**
2. **Sierra Leone**
3. **Chad**
4. **Central African Republic**
5. **Mali**
6. **Niger**
7. **Angola**
8. **Congo, Dem. Rep.**
9. **Burkina Faso**
10. **Guinea Bissau**

Summary

K-Means vs Hierarchical Clustering

K-means clustering :

Countries that are direst need of aid

- **Total 55 countries are in this category**
- **We will consider this now ignoring the countries of the other clusters as we are only concentrating on the countries with poor socio economic factor.**

Hierarchical clustering

Countries that are in the direst need of the aid

- **We have total 158 countries in this category.**
- **There are 103 countries more.**
- **we analyzed and came to conclusion to keep the similar countries as suggested by the k-mean clustering too.**

Conclusion

- We have seen from both methods - (K-Means and Hierarchical clustering) that extra 103 countries are being selected from hierarchical clustering.
- I would choose the final countries from k-means clustering as it gave accurate output than hierarchical clustering.
- I have compared the clusters and visualized from both methods and K-means gave precise information than hierarchical clustering but as both the clustering shows same countries except Nigeria from hierarchical clustering. However will consider k-means countries mostly as it exhibits neat behavior.

Final countries from k-mean clustering + Hierarchical clustering are:

Haiti
Sierra Leone
Chad
Central African Republic
Mali
Niger
Angola
Congo, Dem. Rep
Nigeria
Burkina Faso