HELP INTERNATIONAL NGO ASSIGNMENT

Prakhar Shrivastava

(DS C22 Batch)

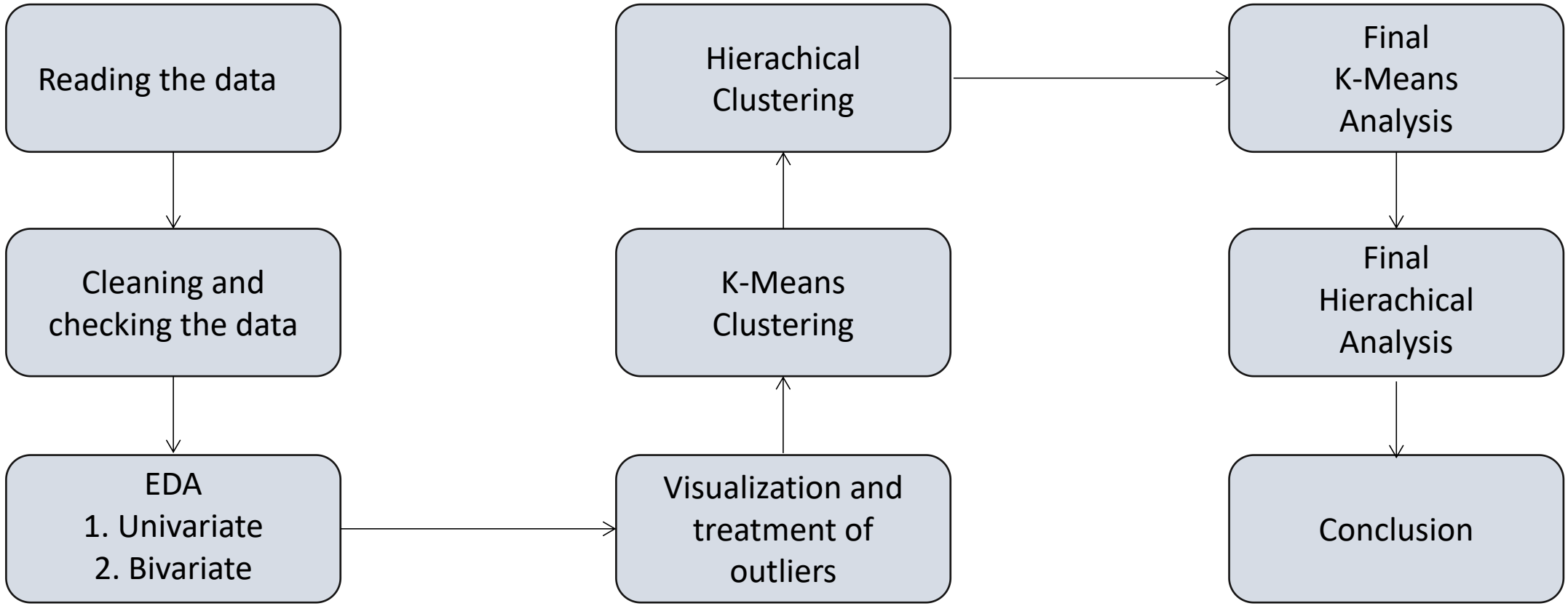
Problem Statement

After the recent funding programmes, an NGO called “HELP International” have raised around $ 10 million. Choosing the countries that are in the direst need of aid will help the CEO of the NGO decide how to use this money strategically and effectively.

Objective

- To categorize the countries using some socioeconomic 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

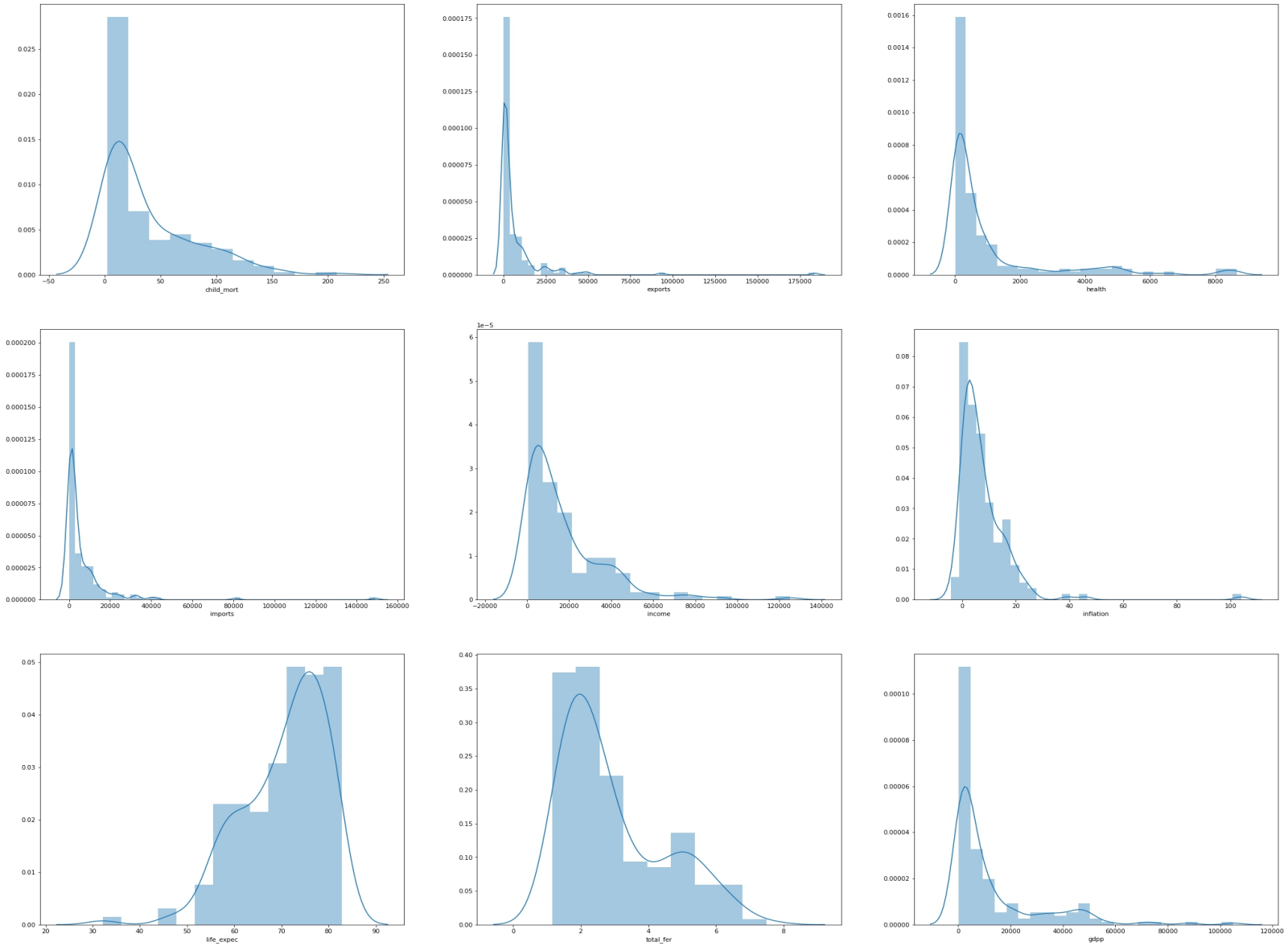
Approach towards the problem



Exploratory Data Analysis

- Univariate

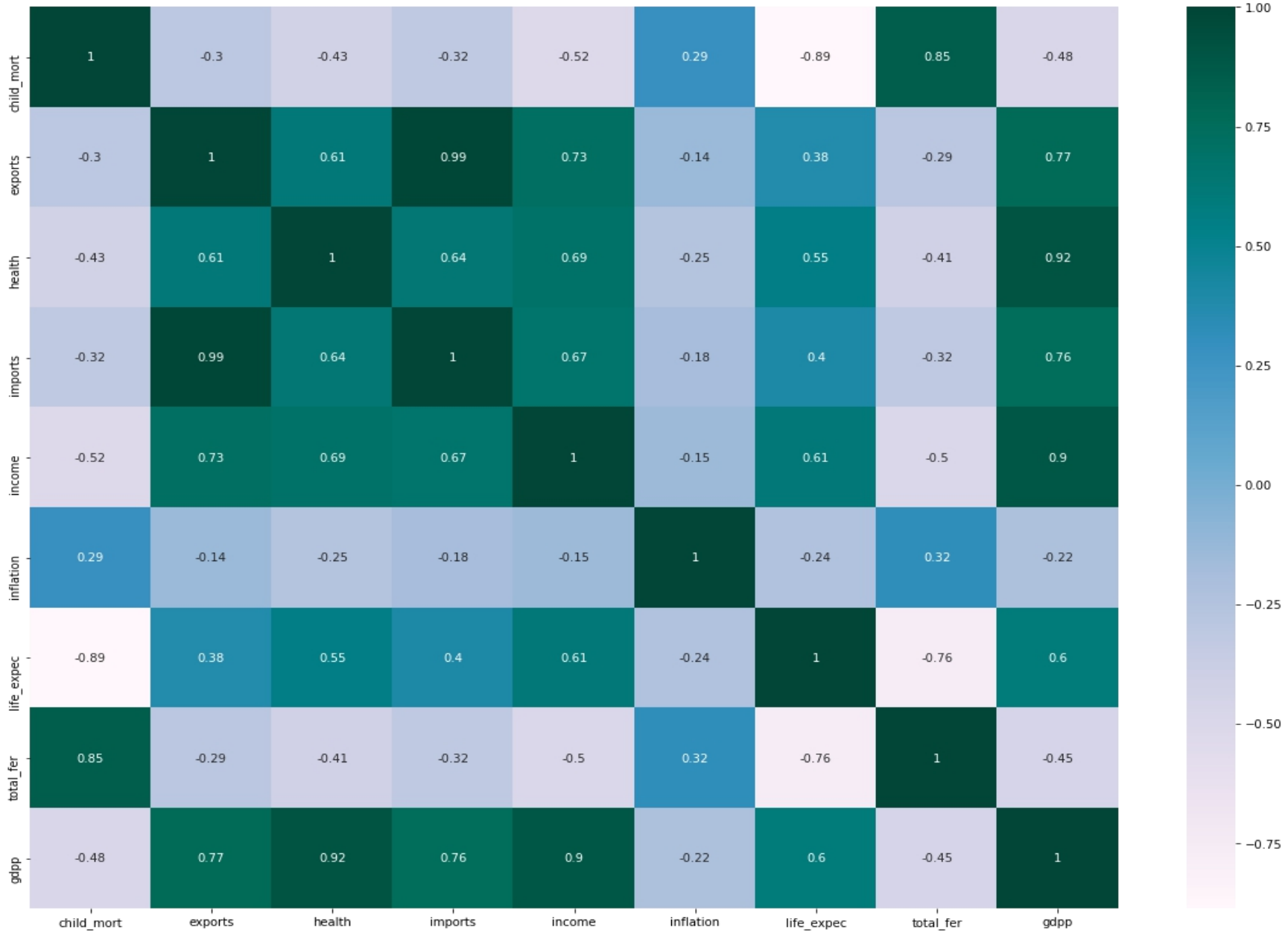
Plotting dist plots to check whether variables are normally distributed or not.



- # Bivariate

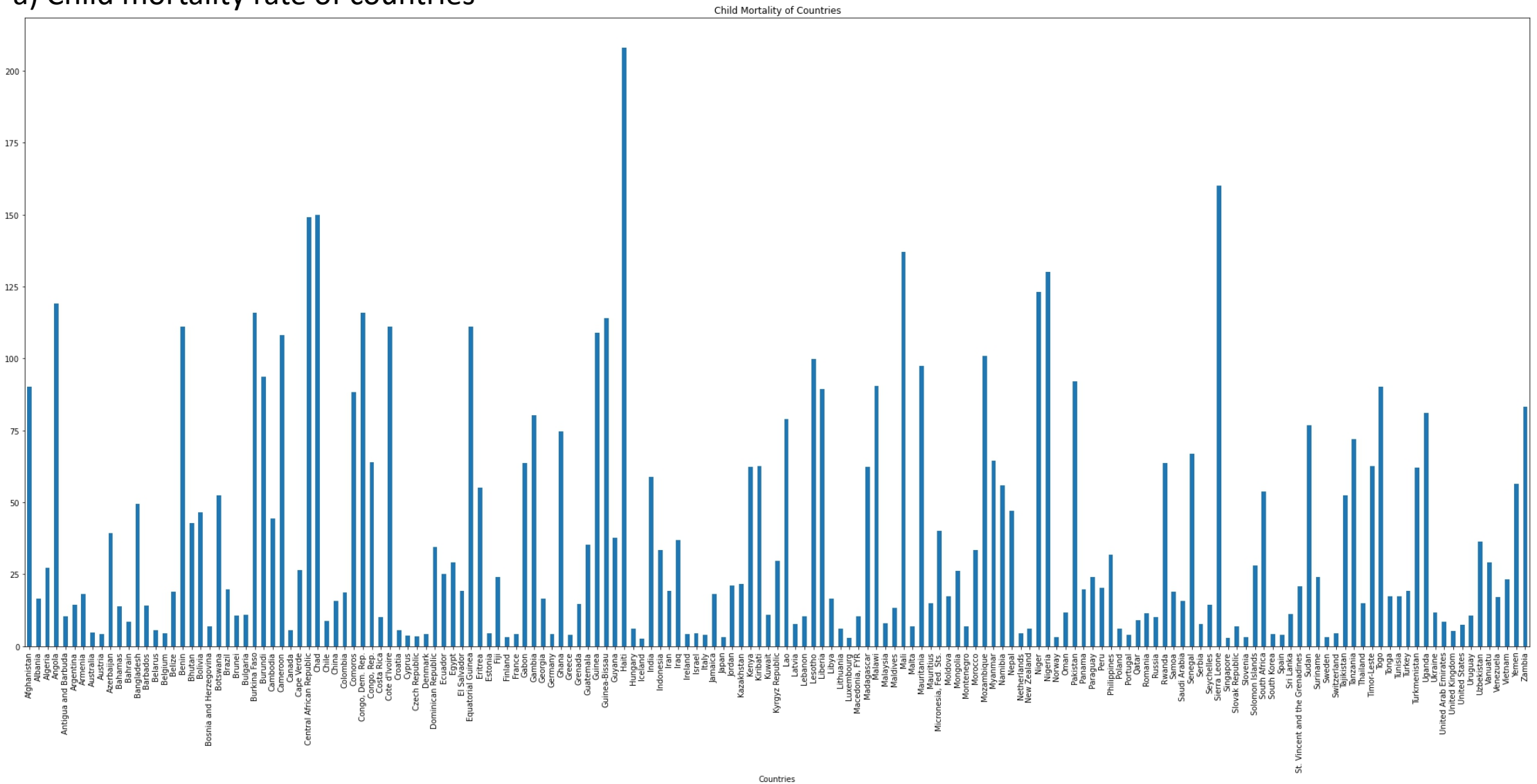
Plotting heat-maps and pair-plots to check the correlation between variables.

- GDPP has high positive correlation with income and health.
- Total fertility rate has high positive correlation with child mortality rate.
- Life expectancy and child mortality have high negative correlation
- Total fertility and life expectancy also have a high negative correlation



- Analysing some important variables under bivariate analysis -

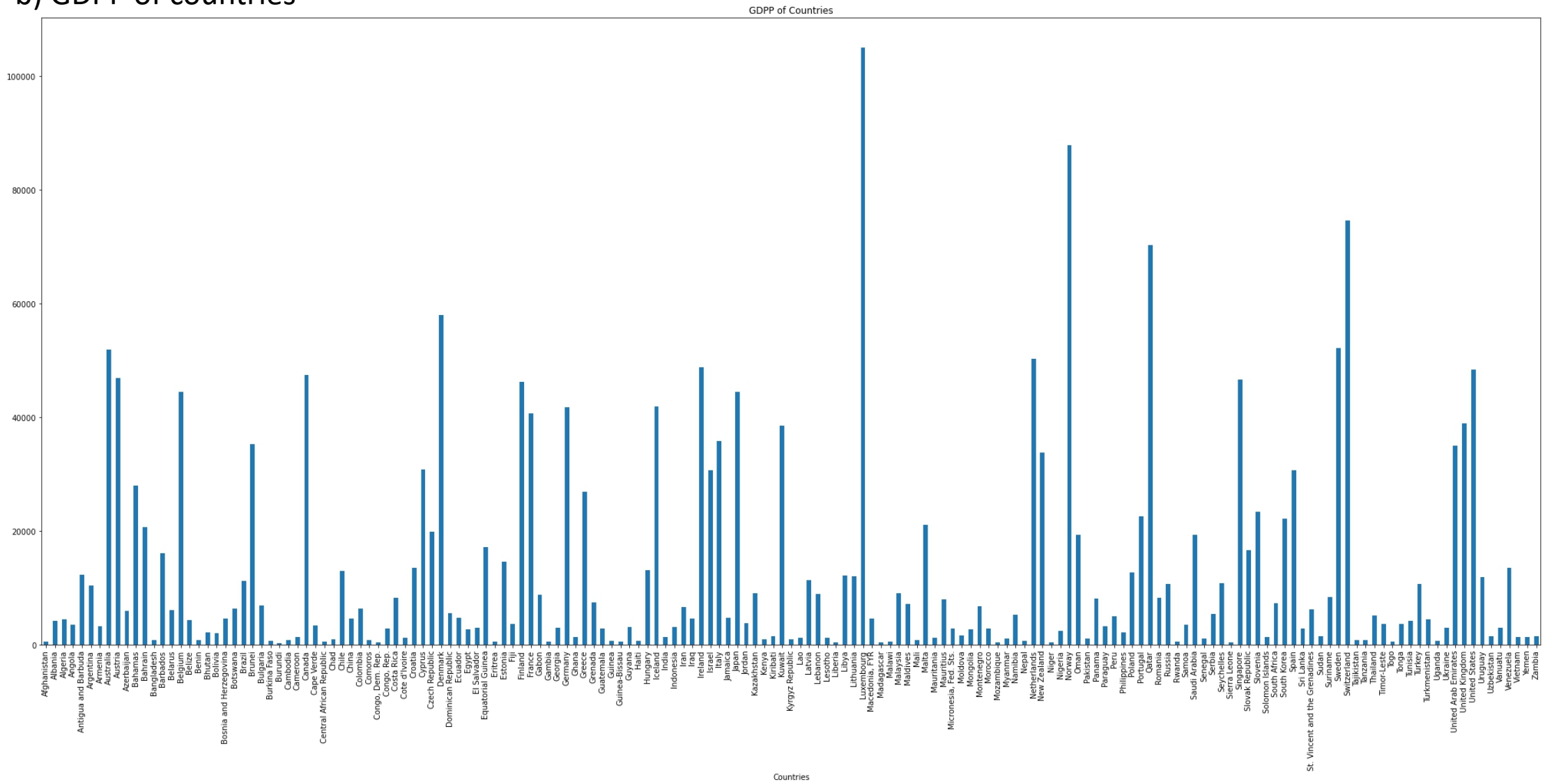
  a) Child mortality rate of countries



As we can see, there are many countries here with appreciable child mortality rates. **Haiti** is the country with highest child mortality rate!

- Analysing some important variables under bivariate analysis -

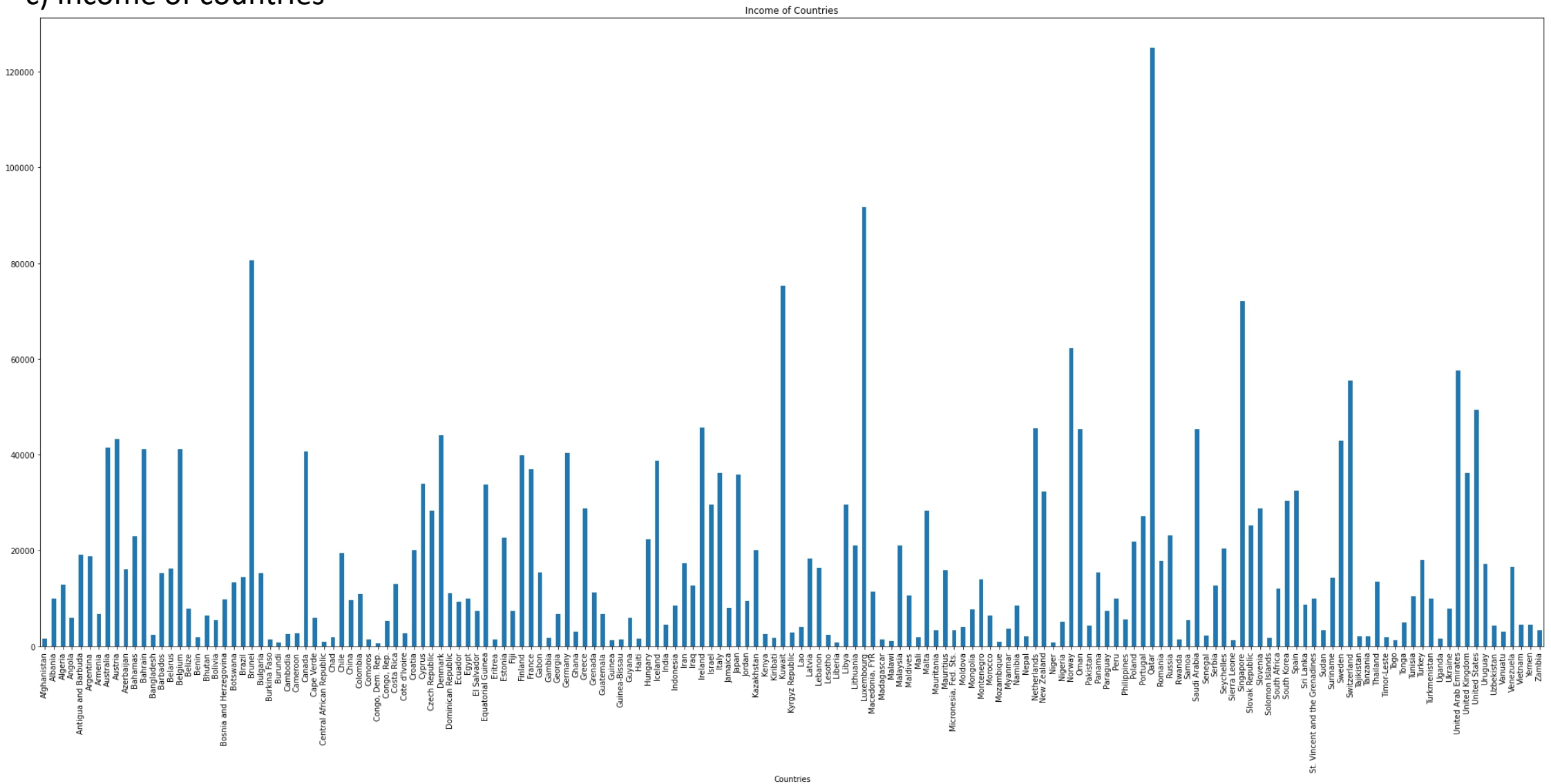
  b) GDPP of countries



As we can see, there are many countries here with considerably low gdpp.
Luxembourg is the country with highest gdpp!

- Analysing some important variables under bivariate analysis -

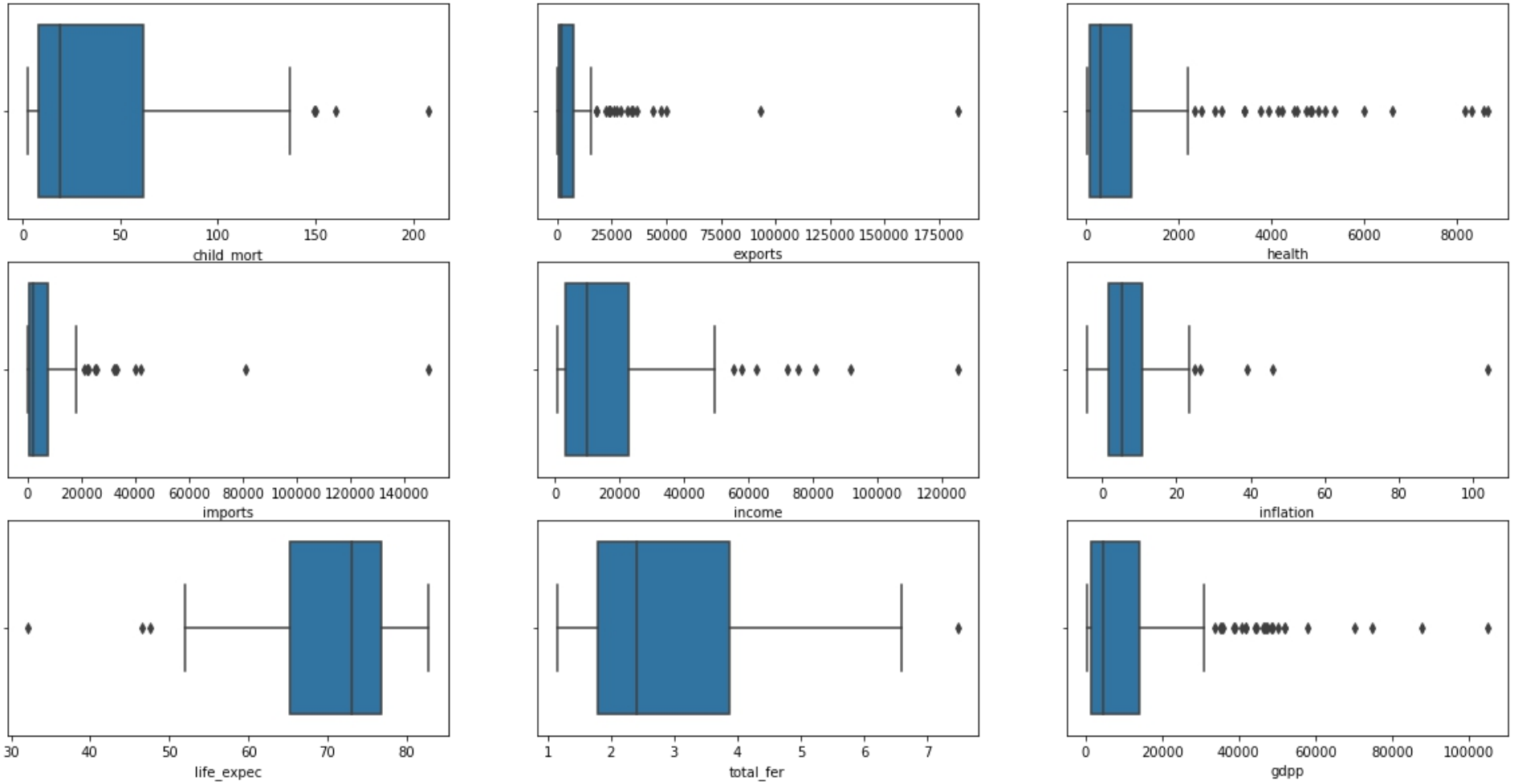
  c) Income of countries



We have a mixed bag of countries with all levels of income.
Qatar is the country with highest income!

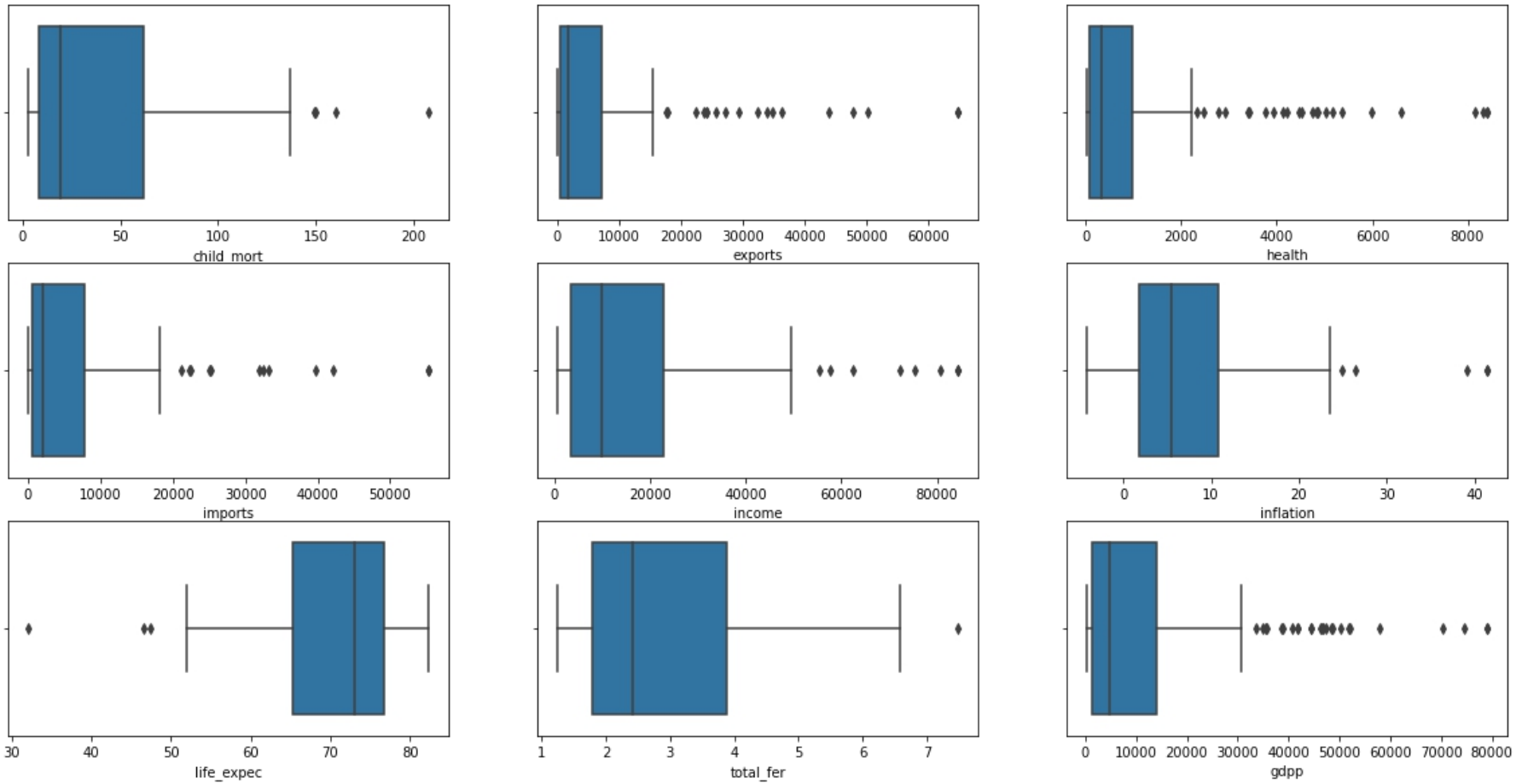
Visualizing and treating outliers

- Visualization



Treatment

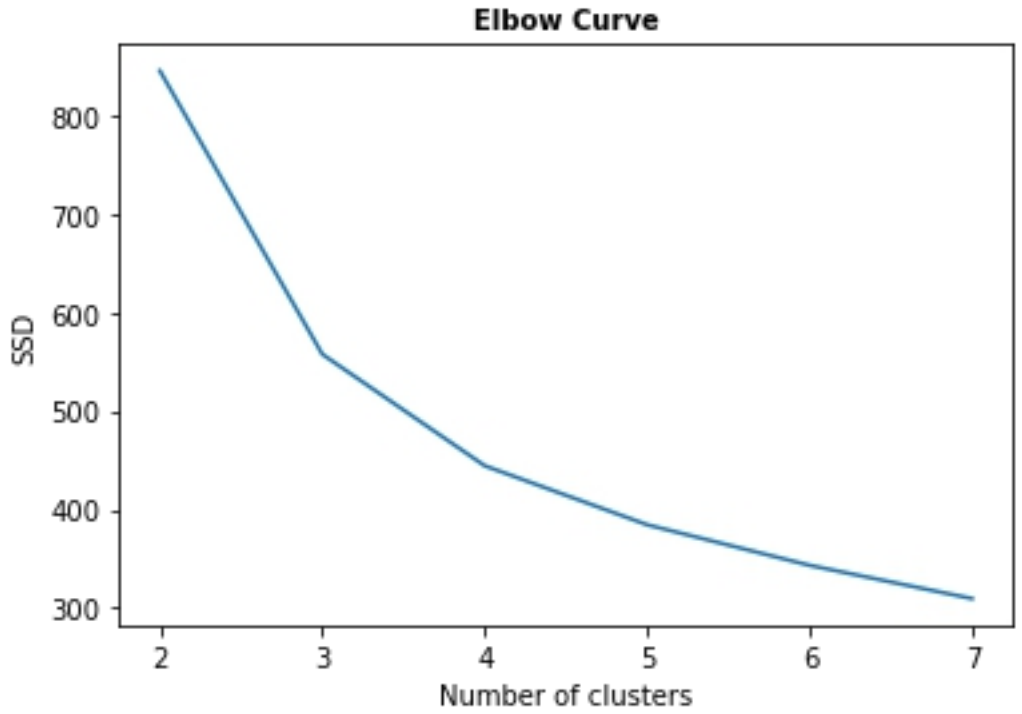
- In this case since we already have very less data, on removing outliers, we might end up losing a lot of critical and useful information.
- Next best method here, is soft capping since there are not many outliers.
- Also, we need to do capping on the upper range on all columns except "child_mort" and "inflation", since we don't want to miss data regarding countries that in need of the aid.
- For "child_mort" and "inflation", we will go with capping of low range outliers.



Outliers post treatment

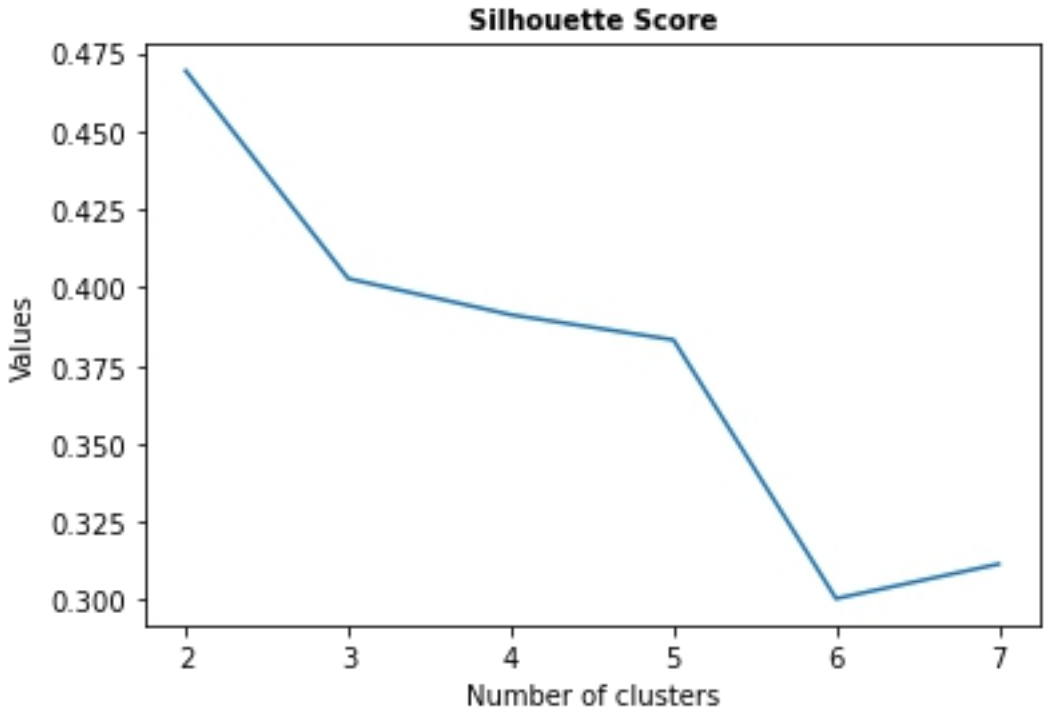
Clustering (K-Means)

- Elbow Curve



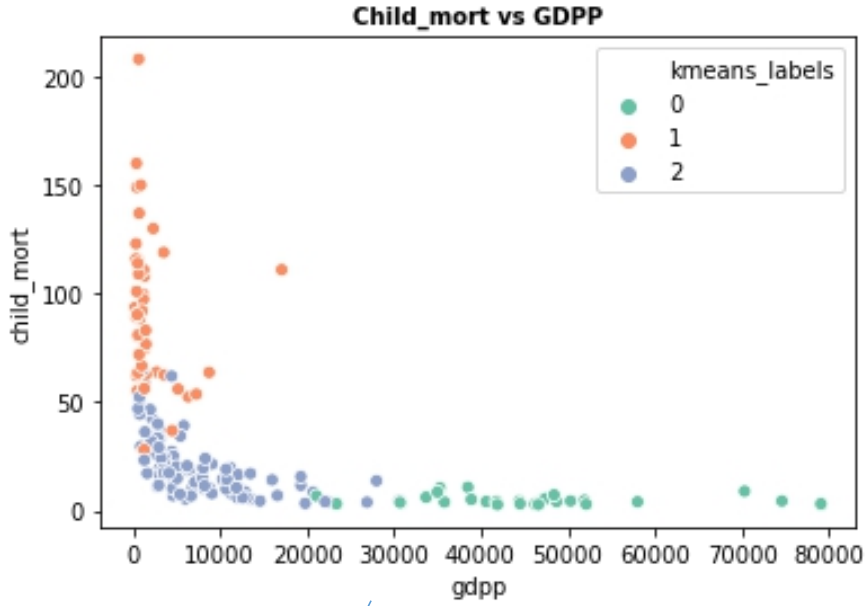
We can see here that on moving from 2 clusters to 3 clusters, there is a steep decrease in SSD, but beyond 3, although the SSD is decreasing, but not at that rate.
Hence, going from 3 cluters to 4 clusters is not fuitful.
We will be going ahead with k = 3, from this graph.

- Silhouette Curve

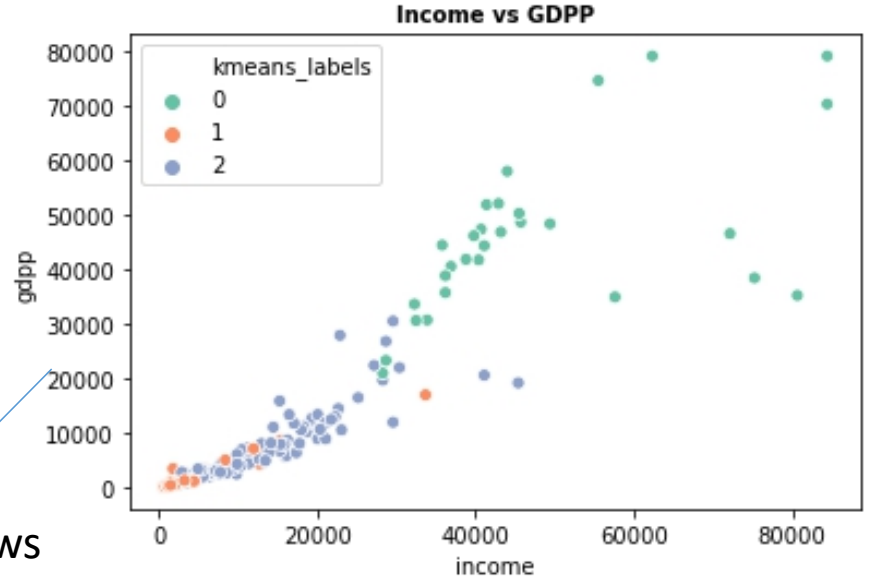


Here, mathematically we can see that num_clusters = 2 is the best because Silhouette score for the same is maximum, but from a business point of view, that might not be the case. Hence, we will go with the next best, i.e. num_clusters = 3.

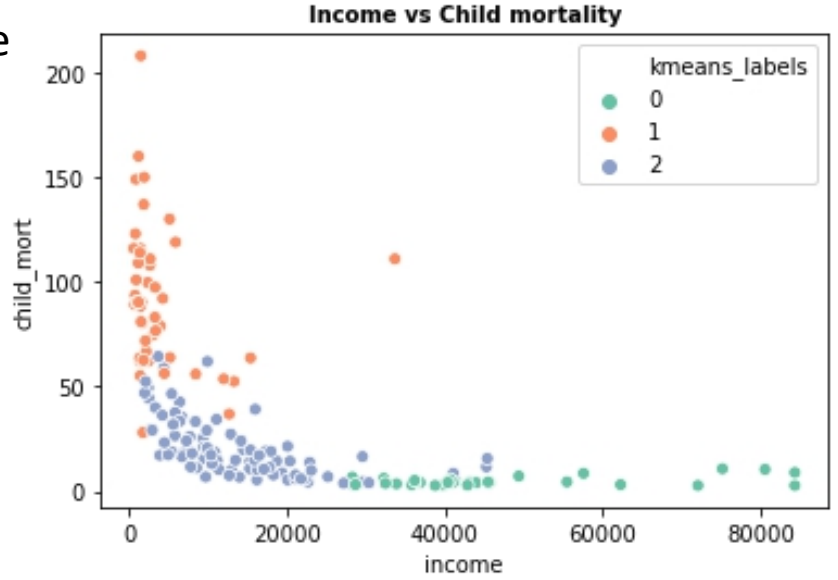
- Visualizing the clusters through scatter plots



It is evident that contries with lesser GDPP tend to have higher child mortality rates.



This graph also shows a pretty much obvious trend, where the GDPP and income increase is going hand in hand.



It is evident that contries with lesser income tend to have higher child mortality rates. This is quite intutive as well!

Top 10 countries in dire need of aid (by K-Means) -

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	kmeans_labels
66	Haiti	208.0	101.2860	45.7442	428.314	1500.0	5.450	32.1	3.33	662.0	1
132	Sierra Leone	160.0	67.0320	52.2690	137.655	1220.0	17.200	55.0	5.20	399.0	1
32	Chad	150.0	330.0960	40.6341	390.195	1930.0	6.390	56.5	6.59	897.0	1
31	Central African Republic	149.0	52.6280	17.7508	118.190	888.0	2.010	47.5	5.21	446.0	1
97	Mali	137.0	161.4240	35.2584	248.508	1870.0	4.370	59.5	6.55	708.0	1
113	Nigeria	130.0	589.4900	118.1310	405.420	5150.0	41.478	60.5	5.84	2330.0	1
112	Niger	123.0	77.2560	17.9568	170.868	814.0	2.550	58.8	7.49	348.0	1
3	Angola	119.0	2199.1900	100.6050	1514.370	5900.0	22.400	60.1	6.16	3530.0	1
37	Congo, Dem. Rep.	116.0	137.2740	26.4194	165.664	609.0	20.800	57.5	6.54	334.0	1
25	Burkina Faso	116.0	110.4000	38.7550	170.200	1430.0	6.810	57.9	5.87	575.0	1

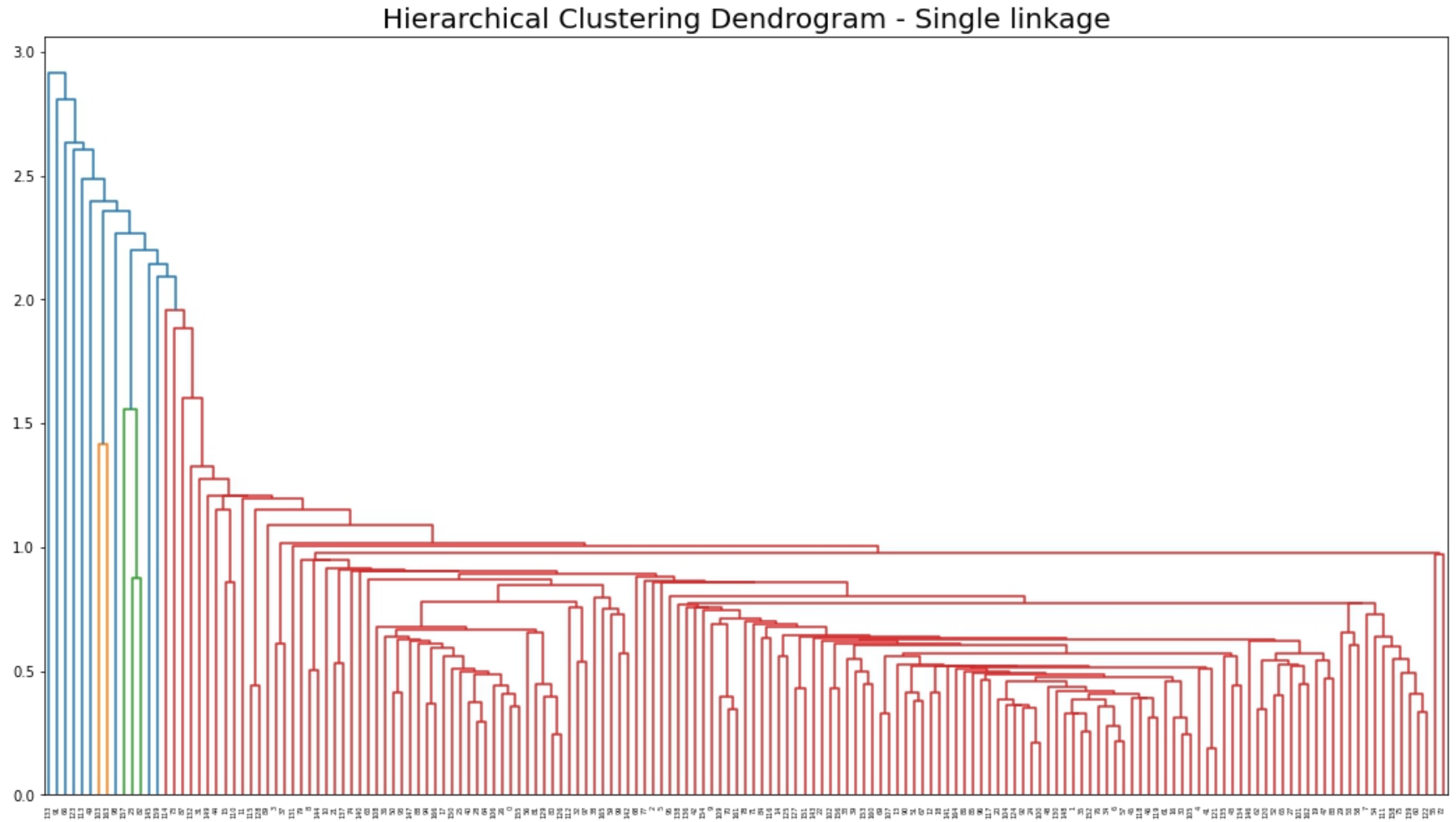
Therefore, according to the K-Means Clustering Algorithm, following 10 countries need aid -

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

Hierarchical Clustering

- Single Linkage

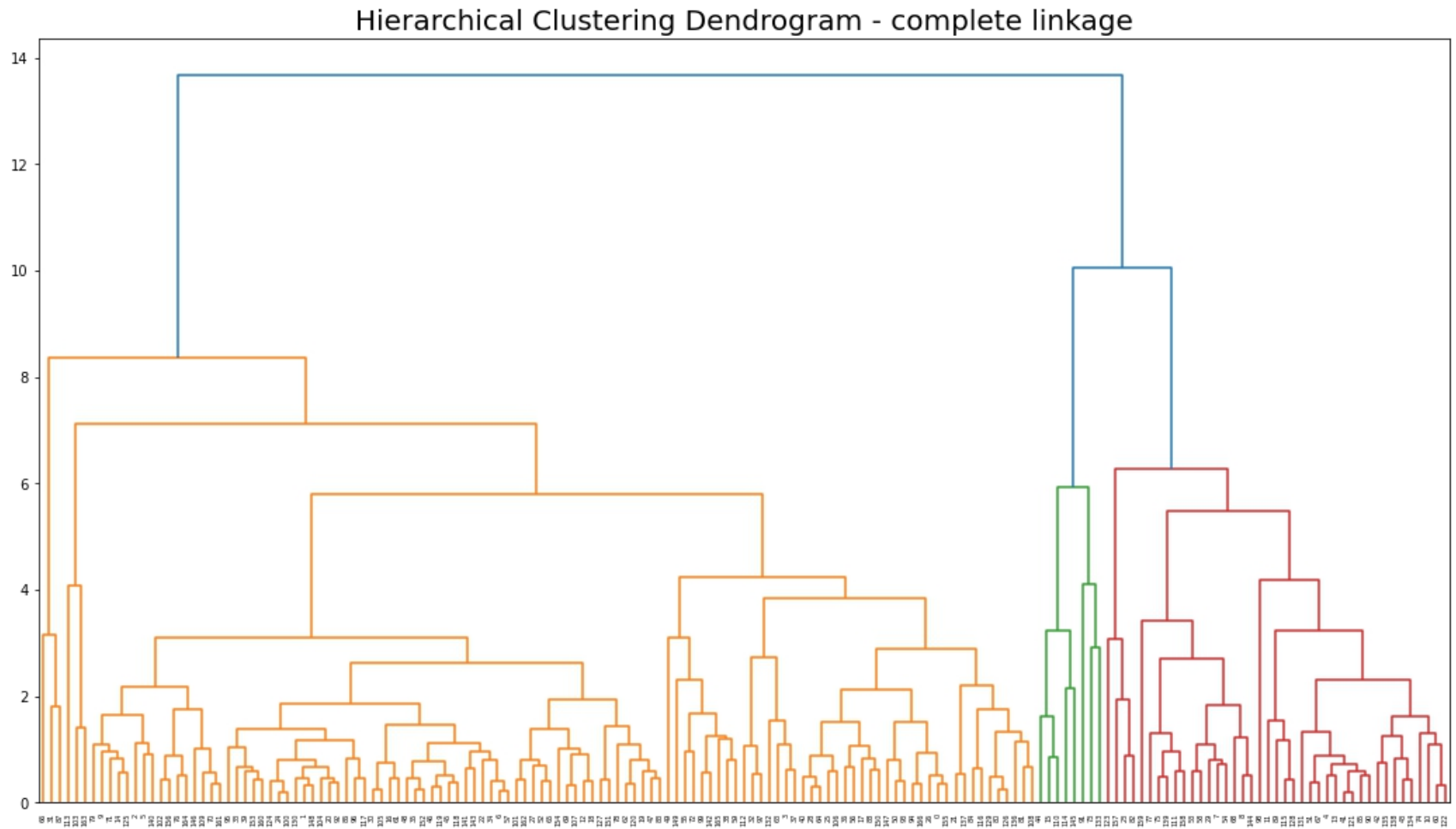
Single linkage Hierarchical clustering was not very clear so we will look into the complete linkage and decide on the cluster for our analysis.



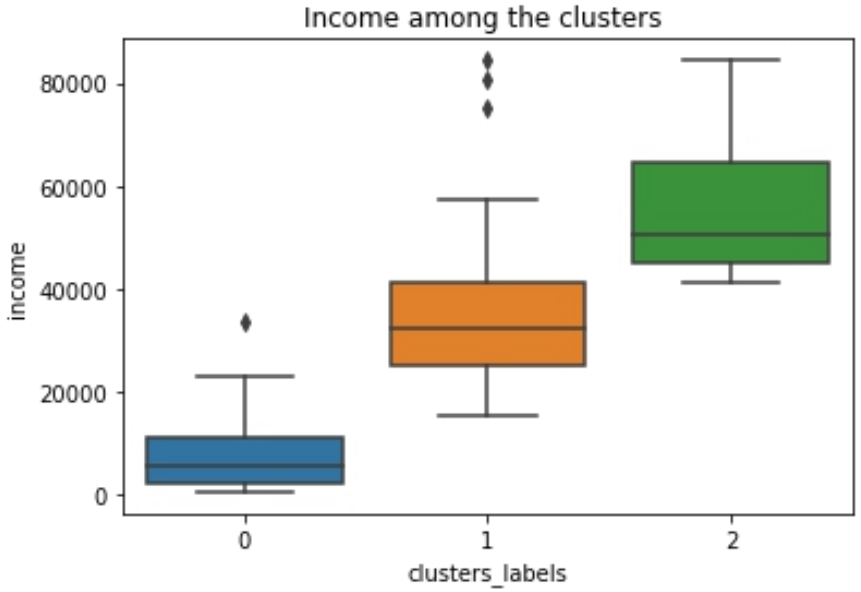
Hierarchical Clustering

- Complete Linkage

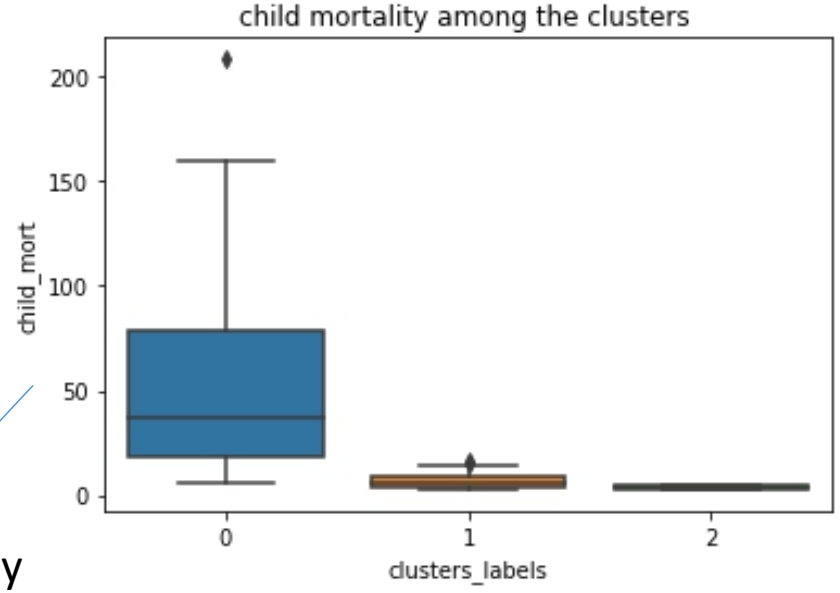
Complete linkage Hierarchical clustering gives us a option to refine the countries into many clusters and analyze.



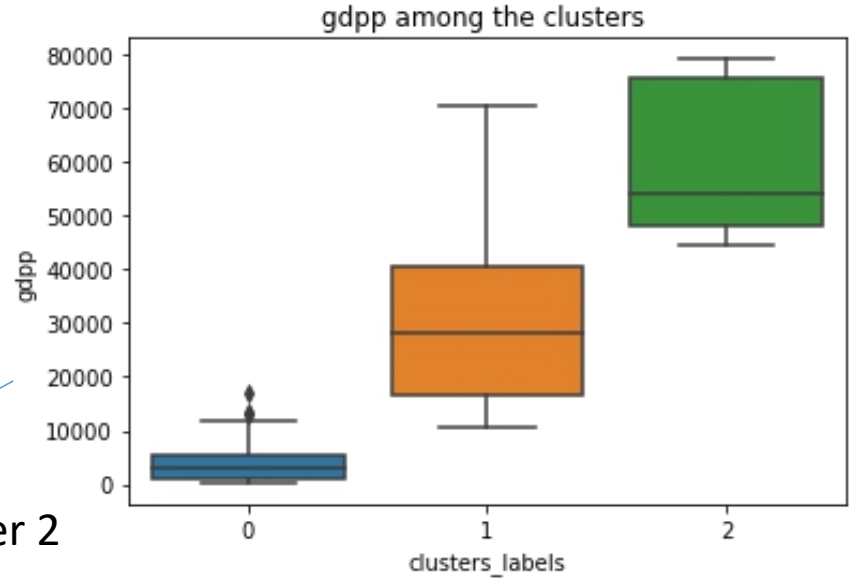
- Analysing the clusters



Countries in cluster 2 have high income level



Here, there are many coutries in cluster 0, that have high child mortality rate



As we can see here, countries in the cluster 2 have high GDPP

Top 10 countries in dire need of aid (by Hierarchical) -

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	clusters_labels
66	Haiti	208.0	101.286	45.7442	428.314	1500.0	5.450	32.1	3.33	662.0	0
132	Sierra Leone	160.0	67.032	52.2690	137.655	1220.0	17.200	55.0	5.20	399.0	0
32	Chad	150.0	330.096	40.6341	390.195	1930.0	6.390	56.5	6.59	897.0	0
31	Central African Republic	149.0	52.628	17.7508	118.190	888.0	2.010	47.5	5.21	446.0	0
97	Mali	137.0	161.424	35.2584	248.508	1870.0	4.370	59.5	6.55	708.0	0
113	Nigeria	130.0	589.490	118.1310	405.420	5150.0	41.478	60.5	5.84	2330.0	0
112	Niger	123.0	77.256	17.9568	170.868	814.0	2.550	58.8	7.49	348.0	0
3	Angola	119.0	2199.190	100.6050	1514.370	5900.0	22.400	60.1	6.16	3530.0	0
37	Congo, Dem. Rep.	116.0	137.274	26.4194	165.664	609.0	20.800	57.5	6.54	334.0	0
25	Burkina Faso	116.0	110.400	38.7550	170.200	1430.0	6.810	57.9	5.87	575.0	0

Therefore, according to the Hierarchical Clustering Algorithm, following 10 countries need aid -

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

Conclusion

After having a close look at the results obtained by K-Means and Hierachical clustering, we can finally name some common countries that appear in both lists.

**Hence the top 10 countries in direst need of aid are- **

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

Thank you!