HELP INTERNATIONAL NGO

CLUSTERING OF THE COUNTRIES

BY: SHRUTI DANDAGI

PROBLEM STATEMENT

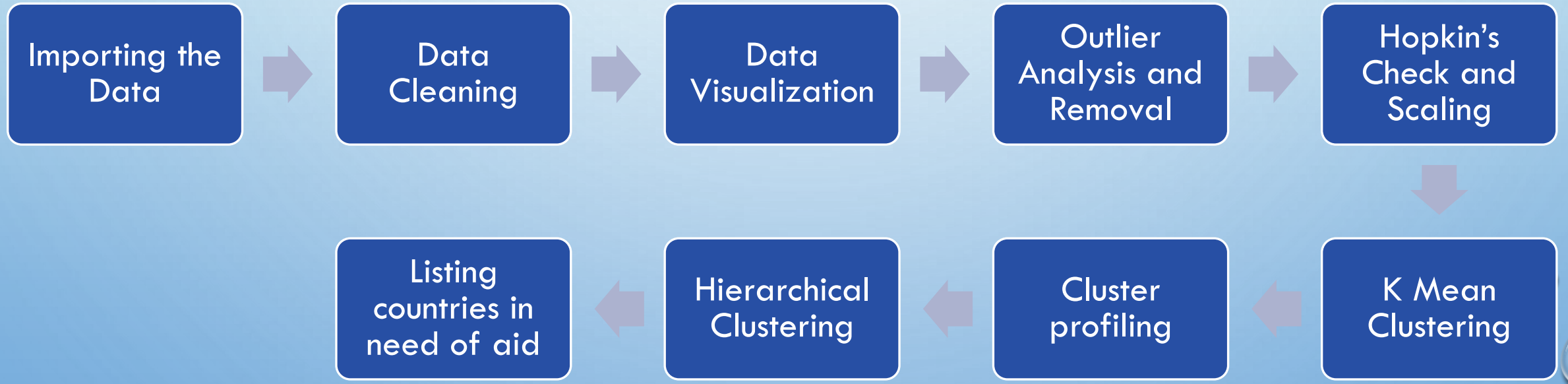
- HELP international is an international humanitarian NGO that is committed to fighting poverty and providing the people of backward countries with basic amenities and relief during the time of disasters and natural calamities. It runs a lot of operational projects from time to time along with advocacy 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.

OBJECTIVE

- Categorise the countries using some socio-economic and health factors that determine the overall development of the country.
- Suggest the countries which needs to be focused more who are in the direst need of aid.

Steps Taken:



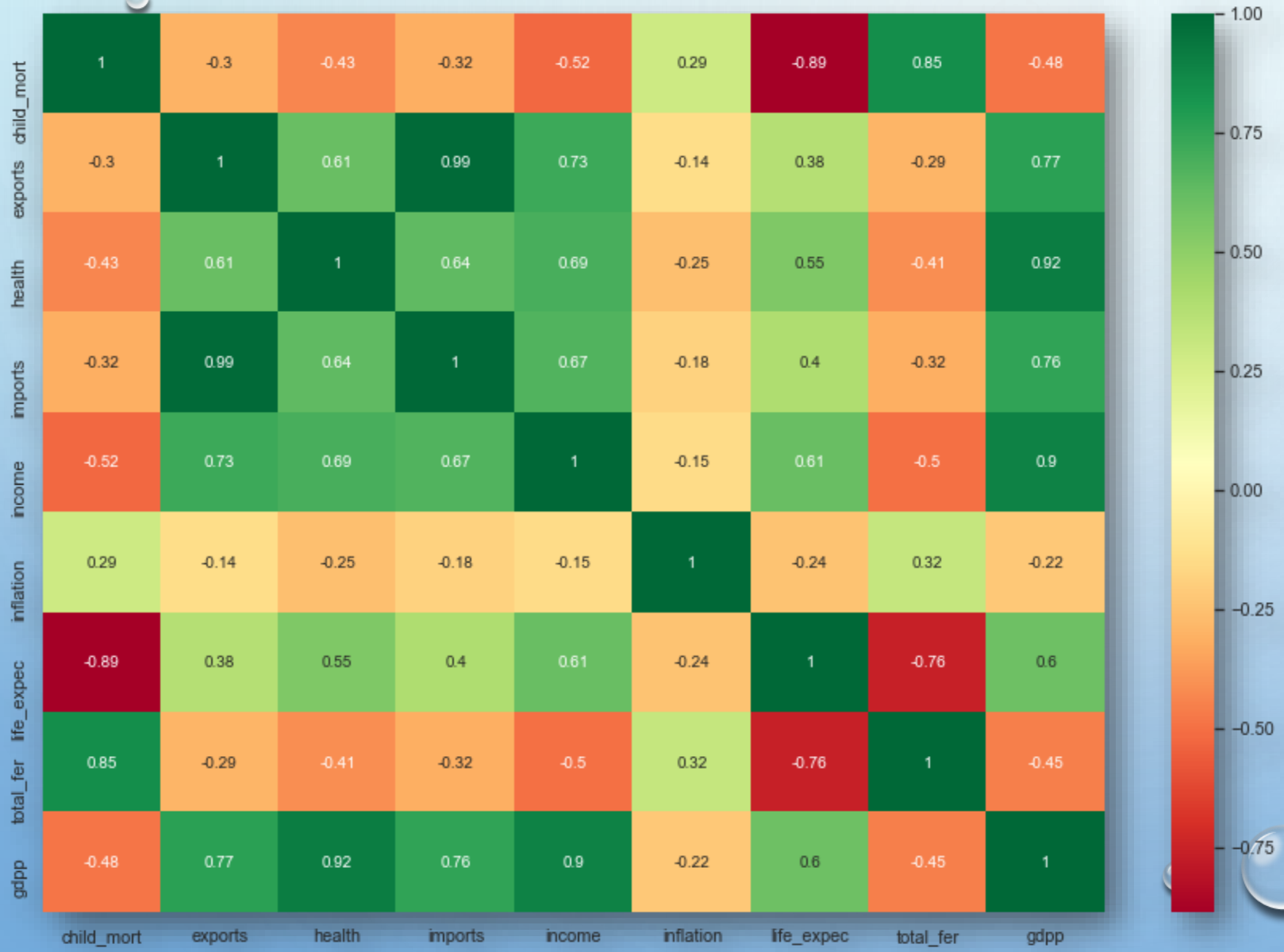
IMPORT, READ AND INSPECT THE DATA

- Data is imported and read from the given country-data csv file.
- The number of rows present are 167 and columns present are 10
- Checked the datatypes of all the columns present
- No missing values in the data
- Summary of all the numeric columns present:

	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
count	167.0000	167.0000	167.0000	167.0000	167.0000	167.0000	167.0000	167.0000	167.0000
mean	38.2701	41.1090	6.8157	46.8902	17144.6886	7.7818	70.5557	2.9480	12964.1557
std	40.3289	27.4120	2.7468	24.2096	19278.0677	10.5707	8.8932	1.5138	18328.7048
min	2.6000	0.1090	1.8100	0.0659	609.0000	-4.2100	32.1000	1.1500	231.0000
25%	8.2500	23.8000	4.9200	30.2000	3355.0000	1.8100	65.3000	1.7950	1330.0000
50%	19.3000	35.0000	6.3200	43.3000	9960.0000	5.3900	73.1000	2.4100	4660.0000
75%	62.1000	51.3500	8.6000	58.7500	22800.0000	10.7500	76.8000	3.8800	14050.0000
max	208.0000	200.0000	17.9000	174.0000	125000.0000	104.0000	82.8000	7.4900	105000.0000

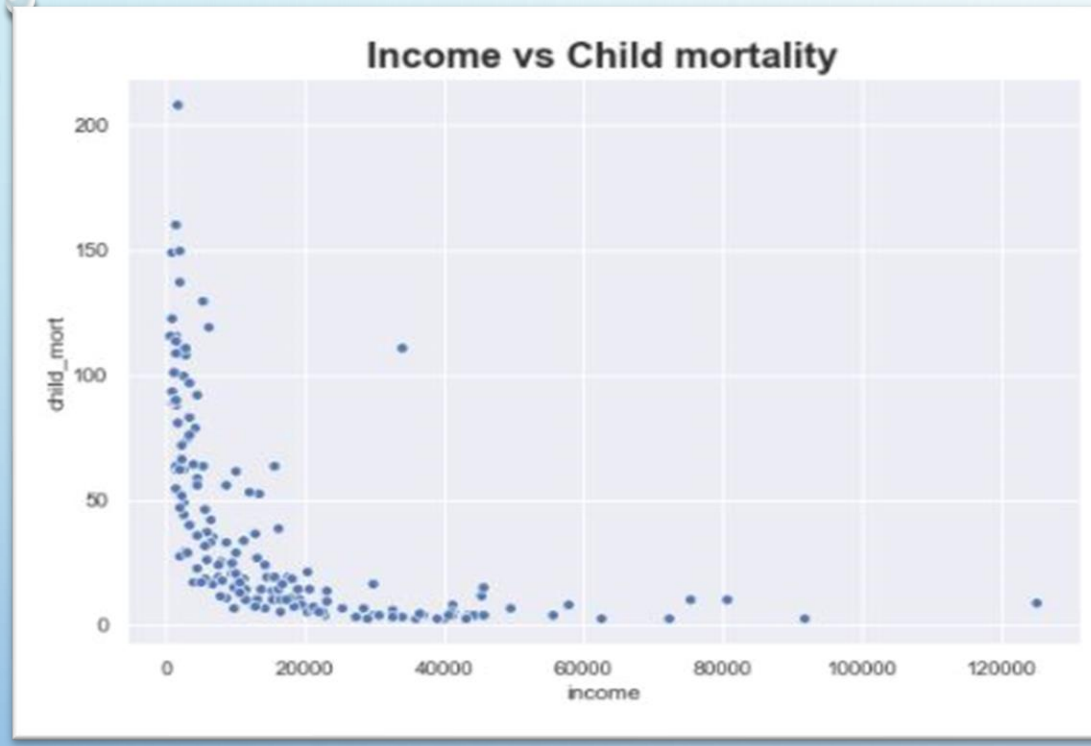
DATA VISUALIZATION

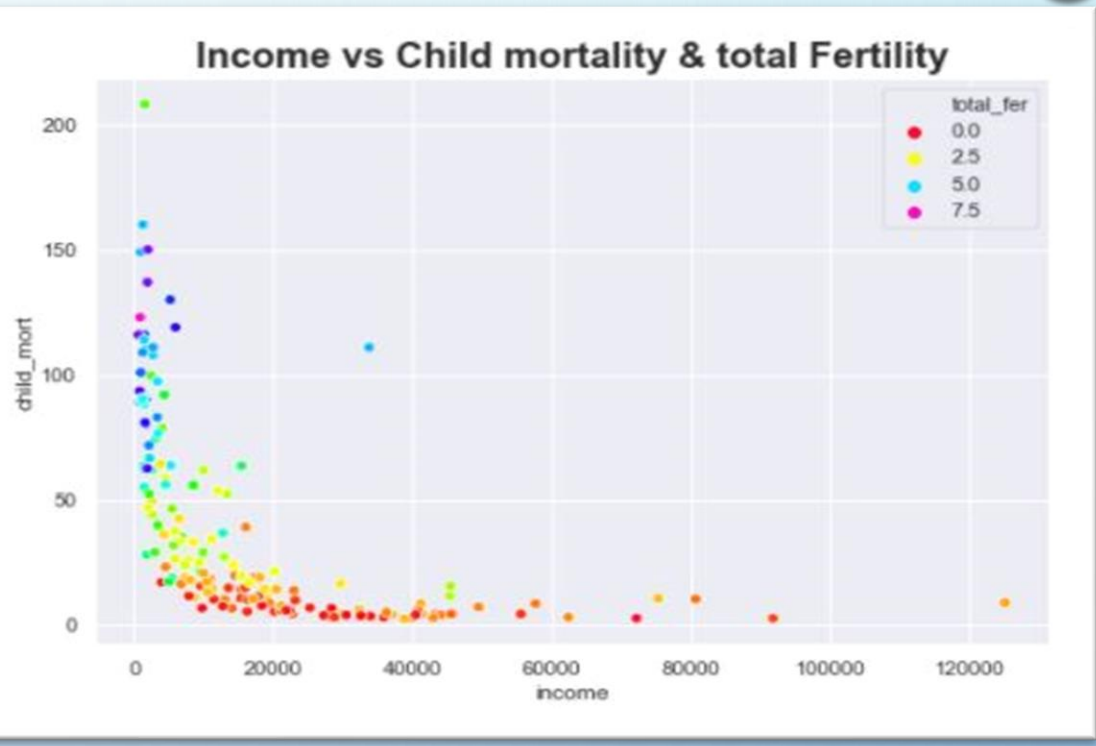
Correlation in the data:



- Child mortality is highly correlated with total fertility with the correlation coefficient of 0.85
- Imports & exports, health & GDP, income & GDP are also well correlated
- Life expectancy has strong negative correlation with total fertility of about -0.76
- Child mortality is also in strong negative correlation with life expectancy

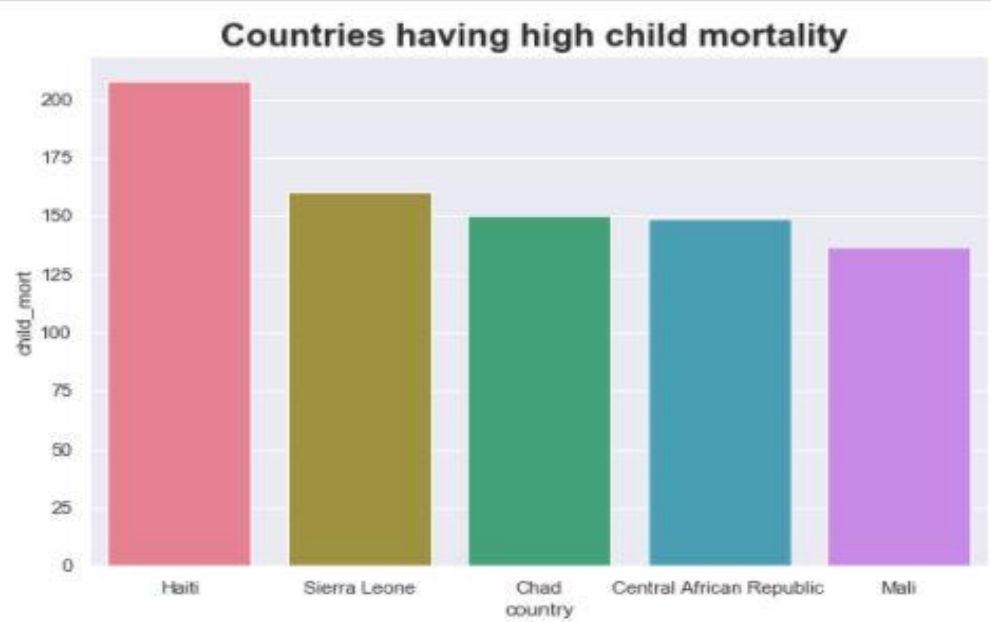
Income , Child Mortality and Total Fertility

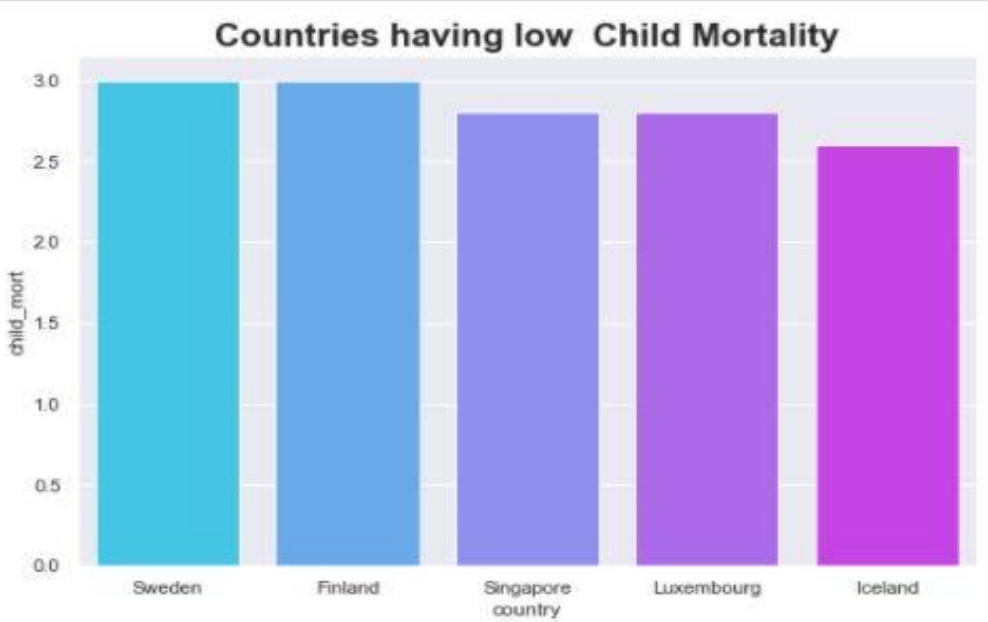




- From the plots above we can see that low income people have high child mortality, which means death of children under age 5 is more, where there is a low income ¶
- Where the income is more we can see there is no mortality
- In the second plot we can see that, high fertility rate for a woman and low income have high child mortality

Country vs child mortality





country	child_mort
Haiti	208.0000
Sierra Leone	160.0000
Chad	150.0000
Central African Republic	149.0000
Mali	137.0000

country	child_mort
Sweden	3.0000
Finland	3.0000
Singapore	2.8000
Luxembourg	2.8000
Iceland	2.6000

- Haiti is a country with highest child mortality of around 208 out of 1000 live births. Next comes Sierra Leone
- Iceland is having very less child mortality of about 2.6.

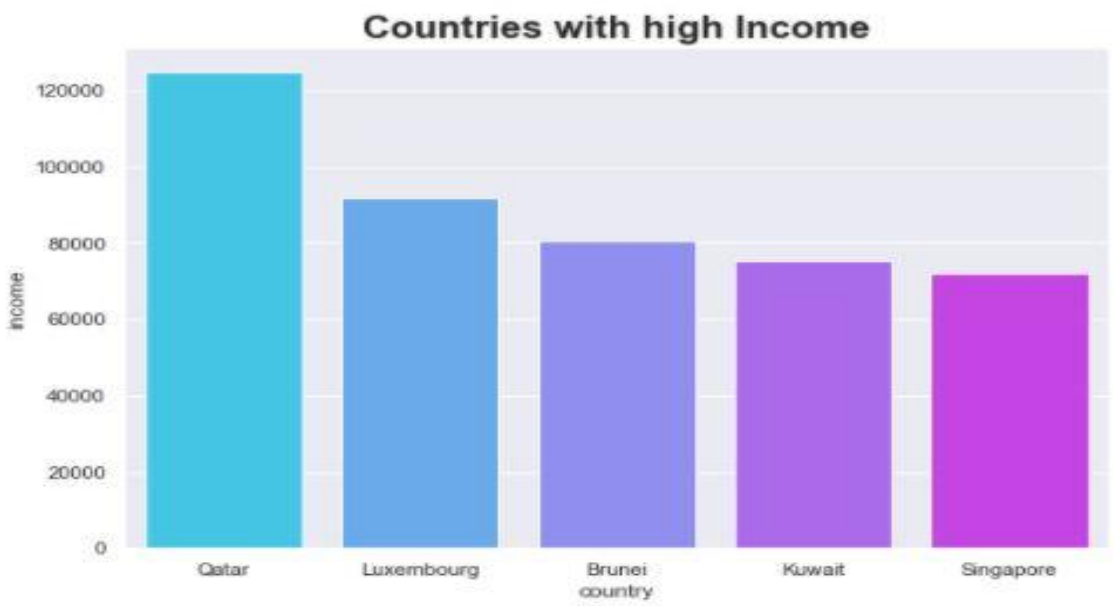
Country vs income

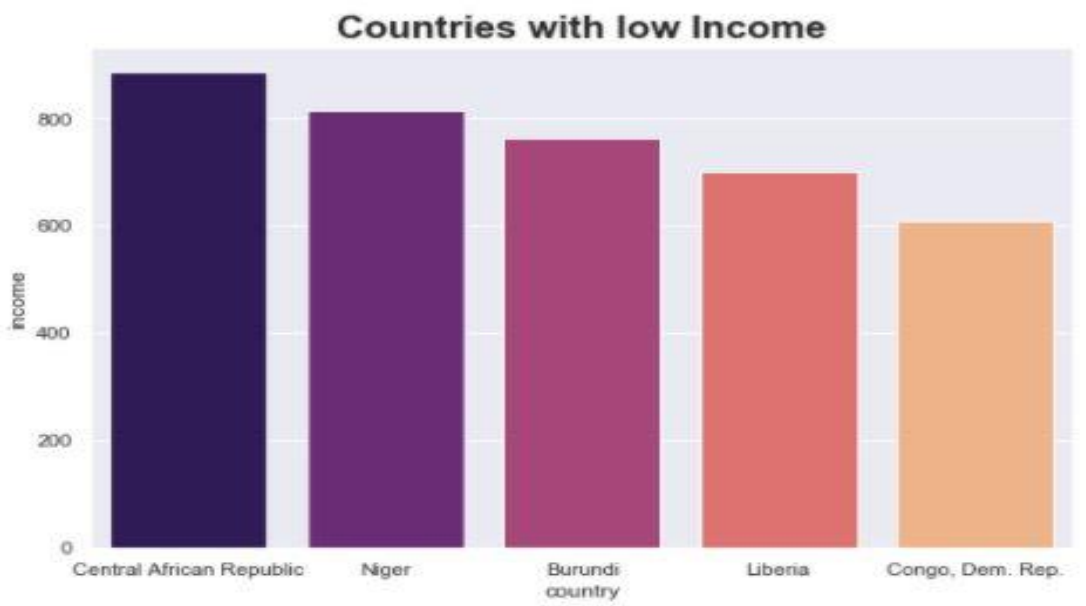
country	income
Qatar	125000
Luxembourg	91700
Brunei	80600
Kuwait	75200
Singapore	72100

country	income
Central African Republic	888
Niger	814
Burundi	764
Liberia	700
Congo, Dem. Rep.	609

- Net income per person is more in Qatar which is 125000
- Luxembourg has net income of around 91700

- Congo, Democratic Republic has less income of about 609
- Liberia has income of about 700





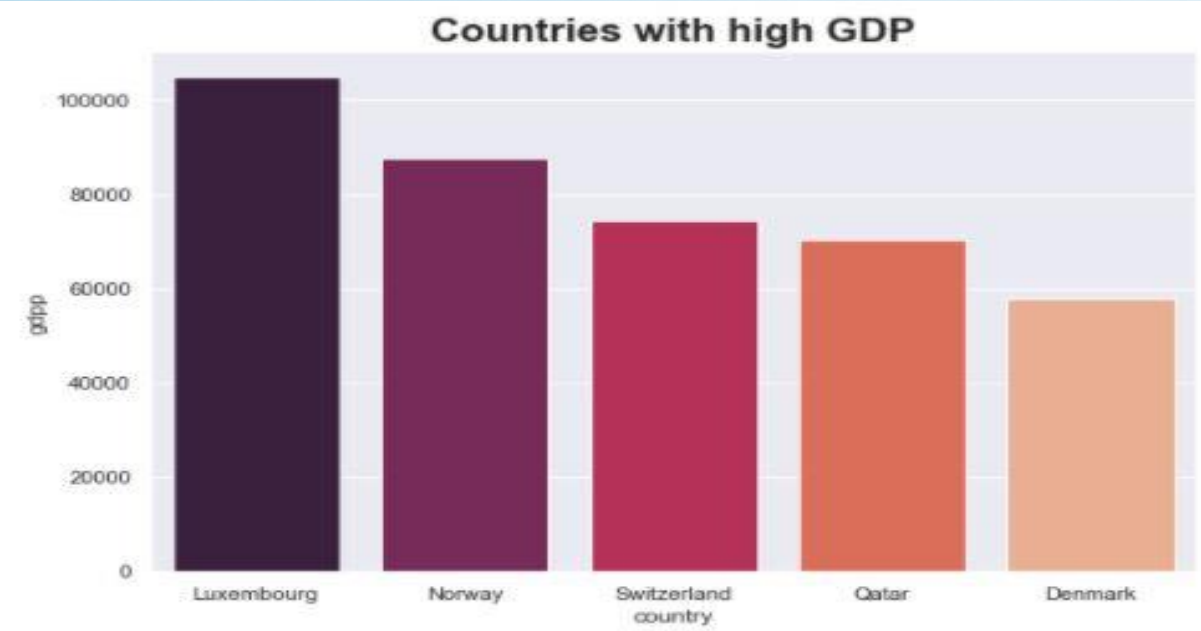
Country vs GDP

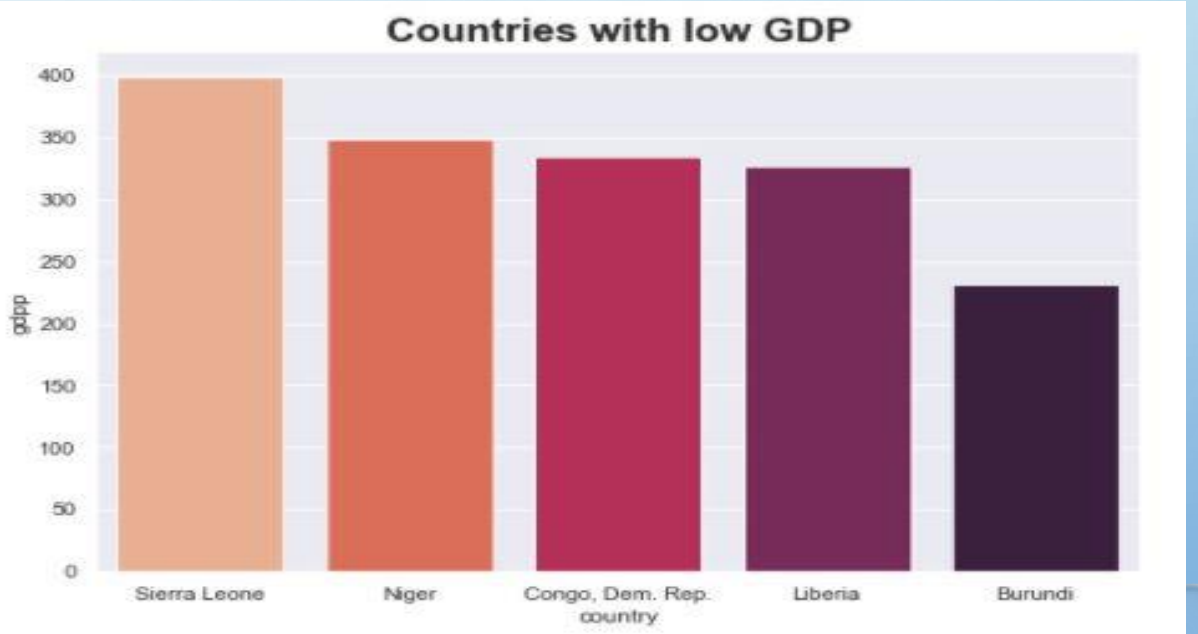
country	gdpp
Luxembourg	105000
Norway	87800
Switzerland	74600
Qatar	70300
Denmark	58000

country	gdpp
Sierra Leone	399
Niger	348
Congo, Dem. Rep.	334
Liberia	327
Burundi	231

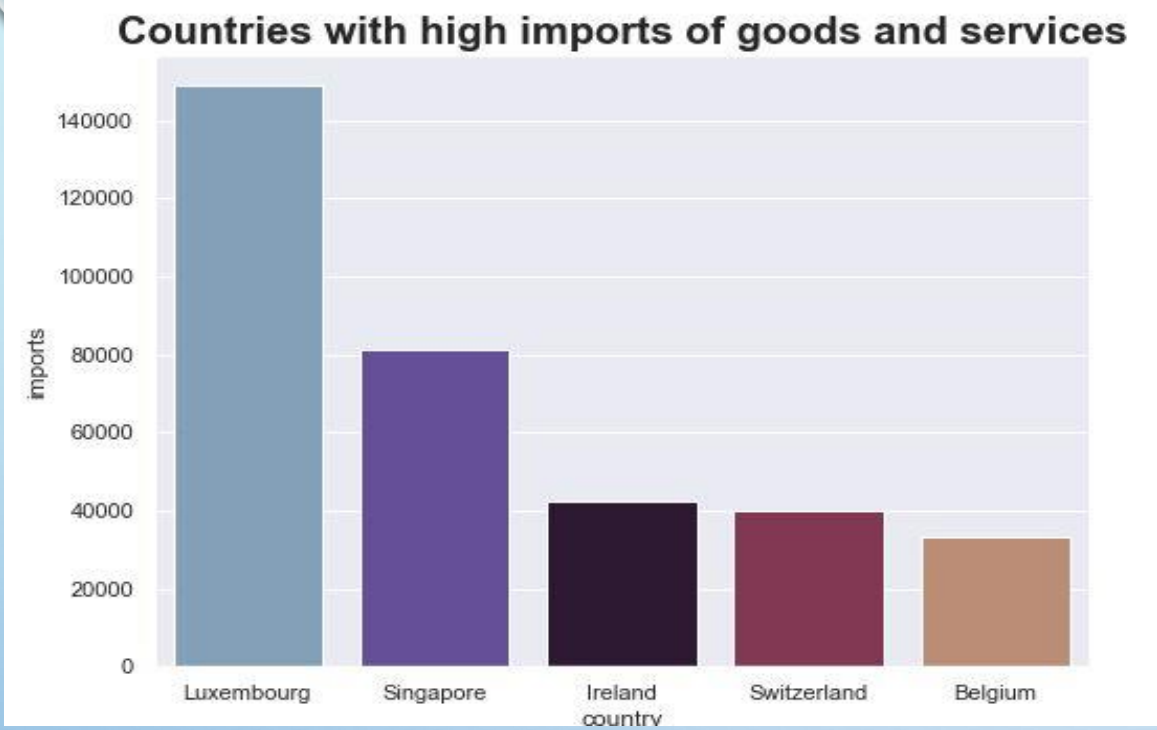
- We can see that Luxembourg has high GDP per capita which is around 105000. Then next comes Norway with GDP of 87800

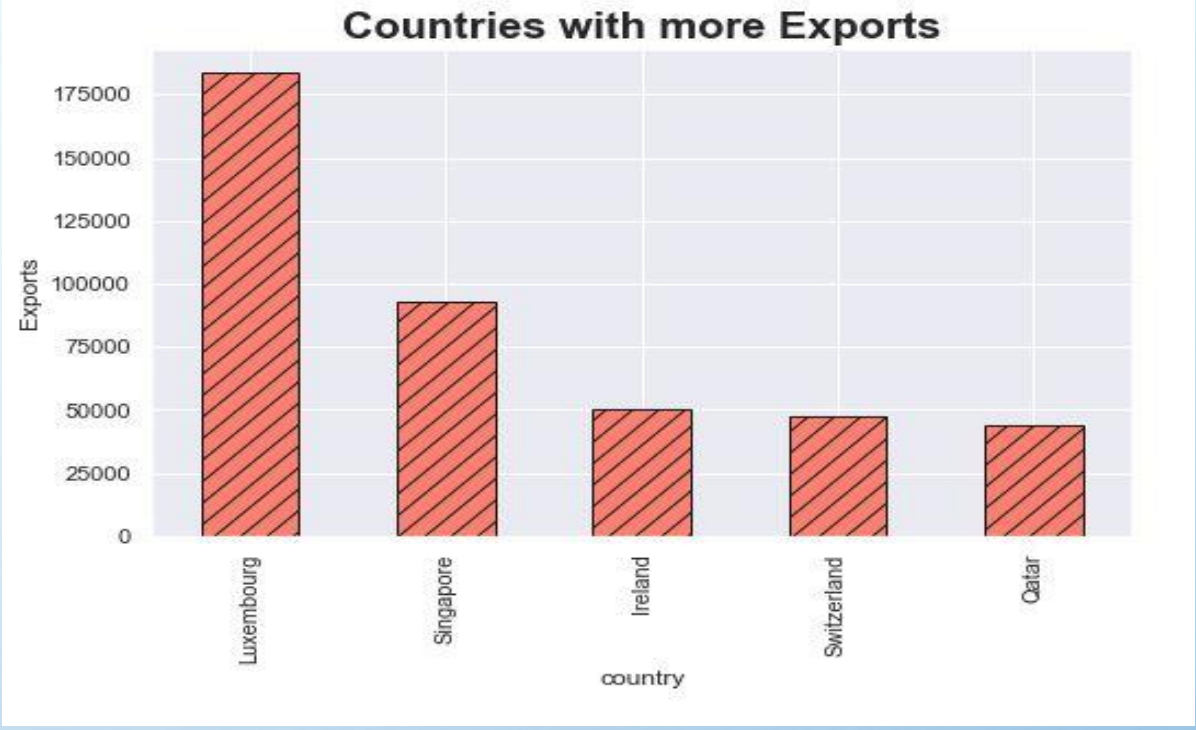
- Burundi has very less GDP per capita of about 231
- Liberia has less GDP of 327





Country vs Imports, Exports





- We can see that Luxembourg has most imports of goods and services of about 149100 per capita

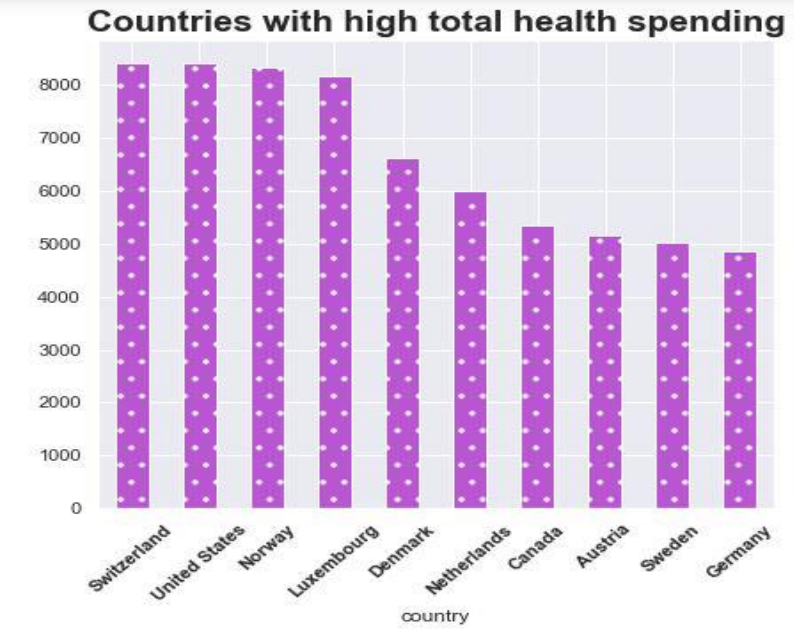
- Myanmar has very less import of goods and services of about 0.6511 per capita

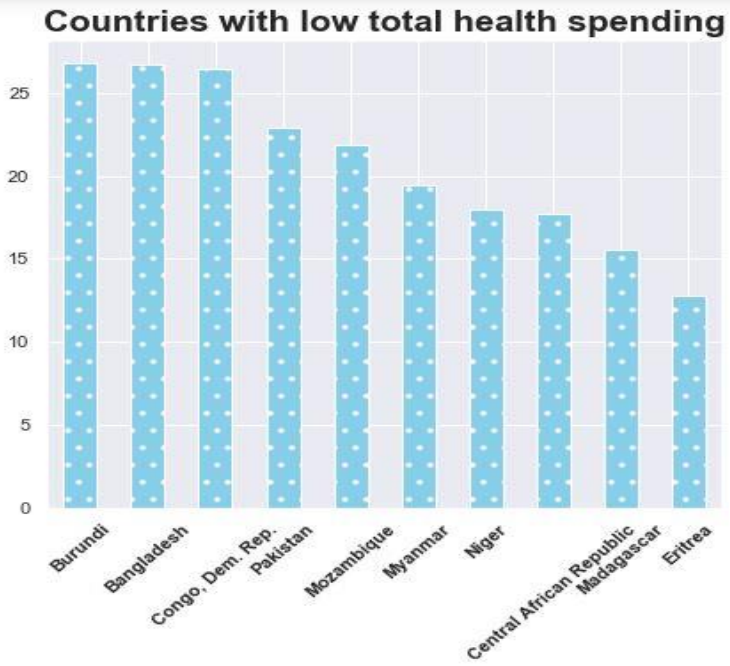
- Luxembourg has more export of goods and services of around 183750, due to which that country has the highest GDP per capita

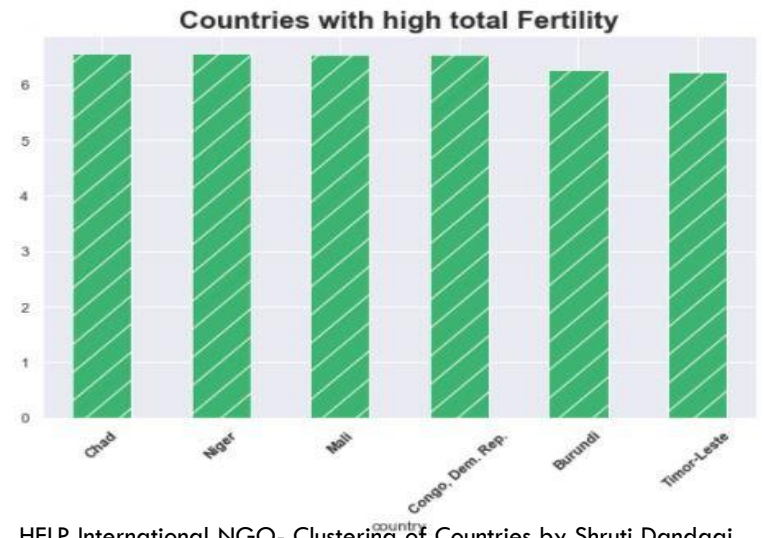
- Myanmar is the country with least exports and next to it is Burundi.

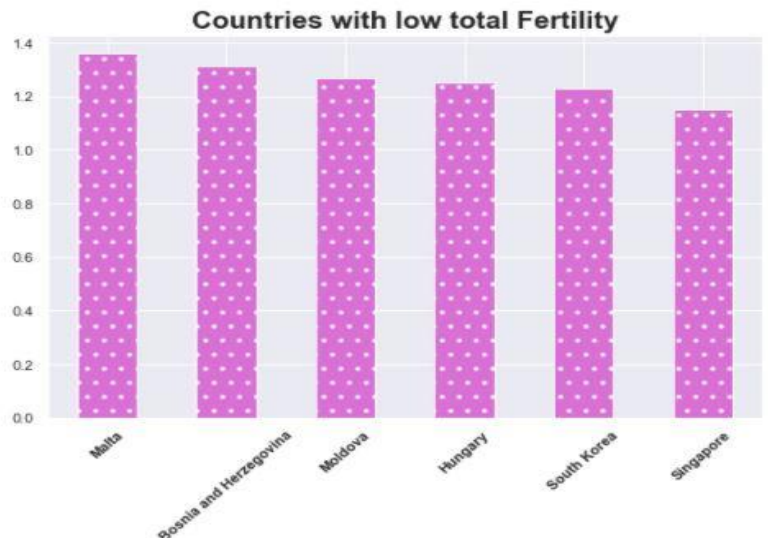
We can make out that same countries with good imports are also good in exports, Countries bad at imports are also bad in exports

Total Health Spending, Inflation, Fertility



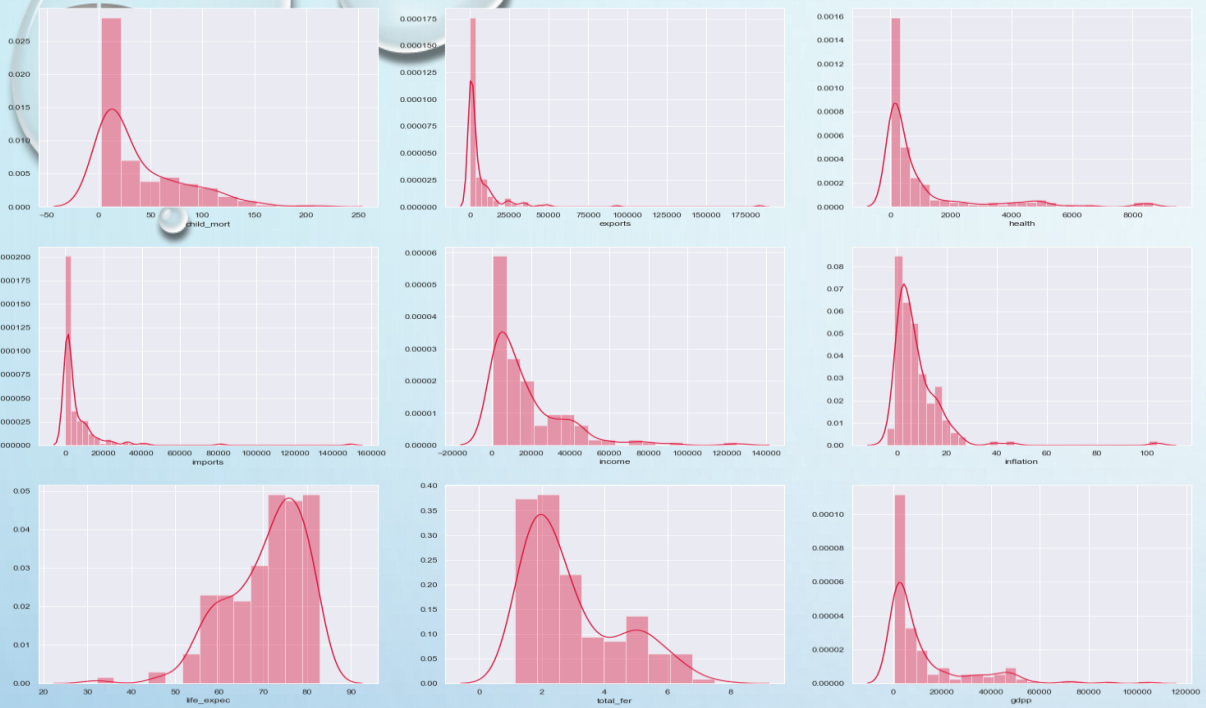






- United States spends more per person on health than comparable countries which is 8663.
- Eritrea, Myanmar, Niger, spends less on health
- Less number of children born to each woman can survive more, As we see Singapore, South Korea have more life expectancy and less total fertility.
- Chad, Mali and Congo Dem Republic have very high fertility and child morality
- Nigeria is having high inflation rate where as Japan, Ireland has very less Inflation rates

Distplot



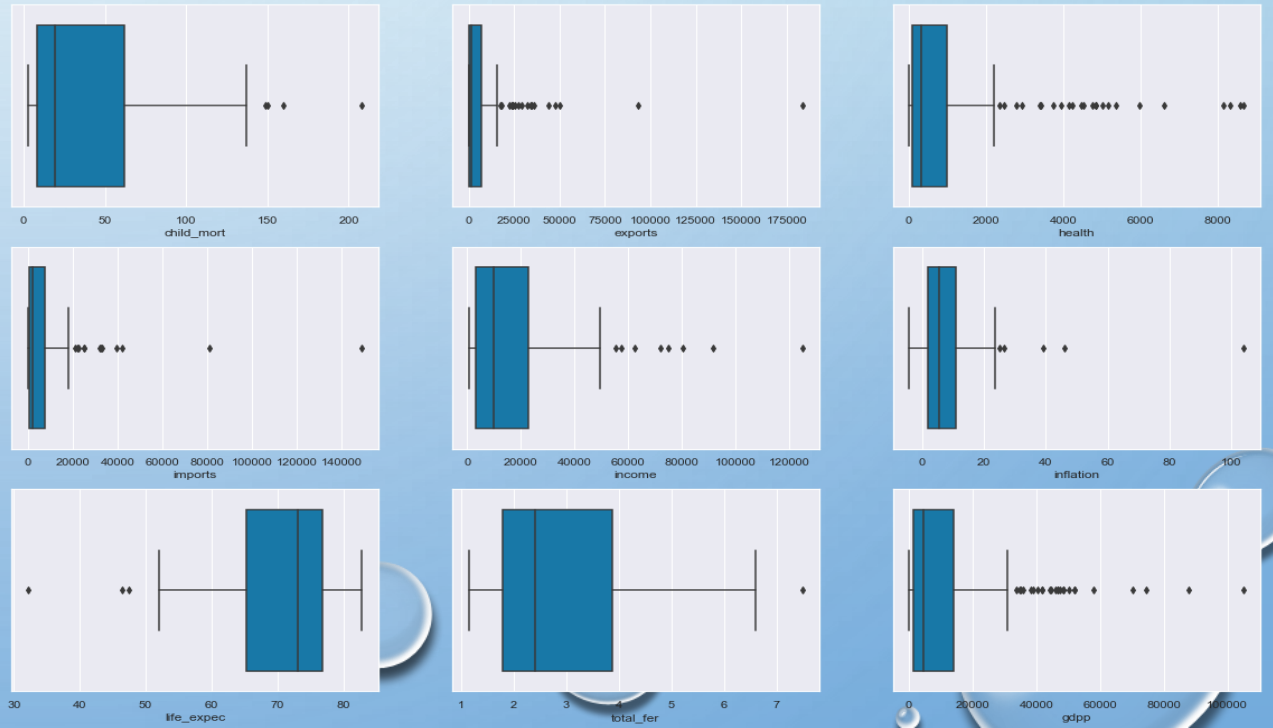
- By looking at the plots we can see that, Child Mortality, Income and GDP are the three columns that have a variation in the data.

- Hence we can consider these three columns for cluster profiling

Boxplot

- There are outliers in the data. We need to treat them as the clustering process is very sensitive to the presence of outliers in the data.

- Outliers are treated by capping them.



CLUSTERING

- Clustering is an unsupervised learning technique, where we try to find patterns based on similarities in the data.
- Two most commonly used types of clustering algorithms - K-Means Clustering and Hierarchical Clustering,
- The Hopkins statistic, is a statistic which gives a value which indicates the cluster tendency, in other words: how well the data can be clustered.
- Performed Hopkins check , and got the value between {0.8, ..., 0.99}, so the dataset has a high tendency to cluster.
- Scaling is performed on the data using Standard scaler.

K-mean Clustering Algorithm Steps:

1. We choose one center as one of the data points at random.

2. For each data point Xi, We compute the distance between Xi and the nearest center that had already been chosen.

3. Now, we choose the next cluster center using the weighted probability distribution where a point X is chosen with probability proportional to d(X)2 .

4. Repeat Steps 2 and 3 until K centers have been chosen.

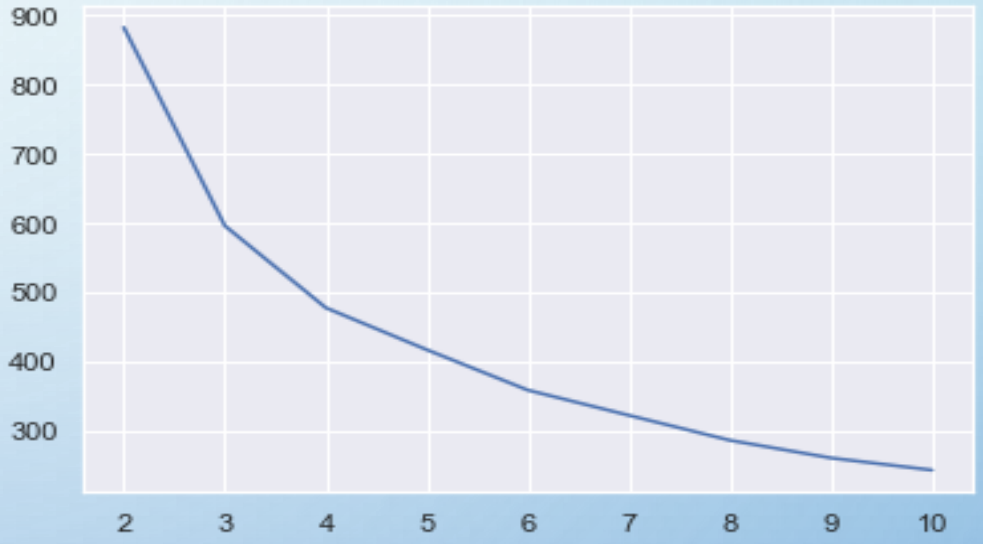
K MEANS CLUSTERING

- To choose the value of k, there are two methods,

1. Silhouette Analysis

From the plot of silhouette score above, We can see that we have the maximum at 2, and next maximum is at 3. So we can go with k= 3.

2. Elbow curve-ssd

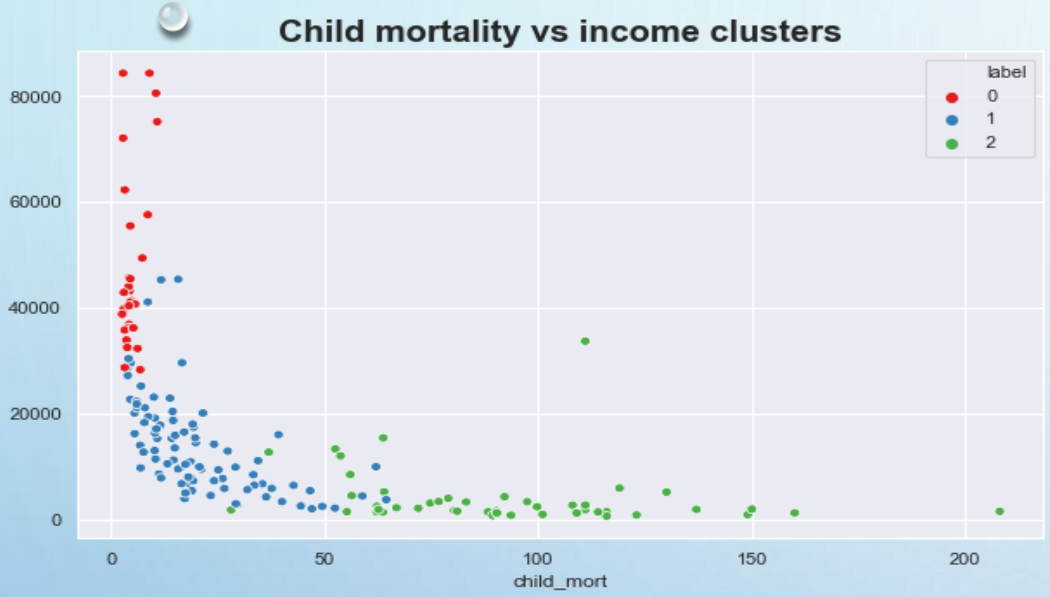


•From the elbow curve above we can see that we have a breakpoint at 3 and one breakpoint at 4.

•From both the plots above, We can go with the lower value for k, which is 3 here. So k=3

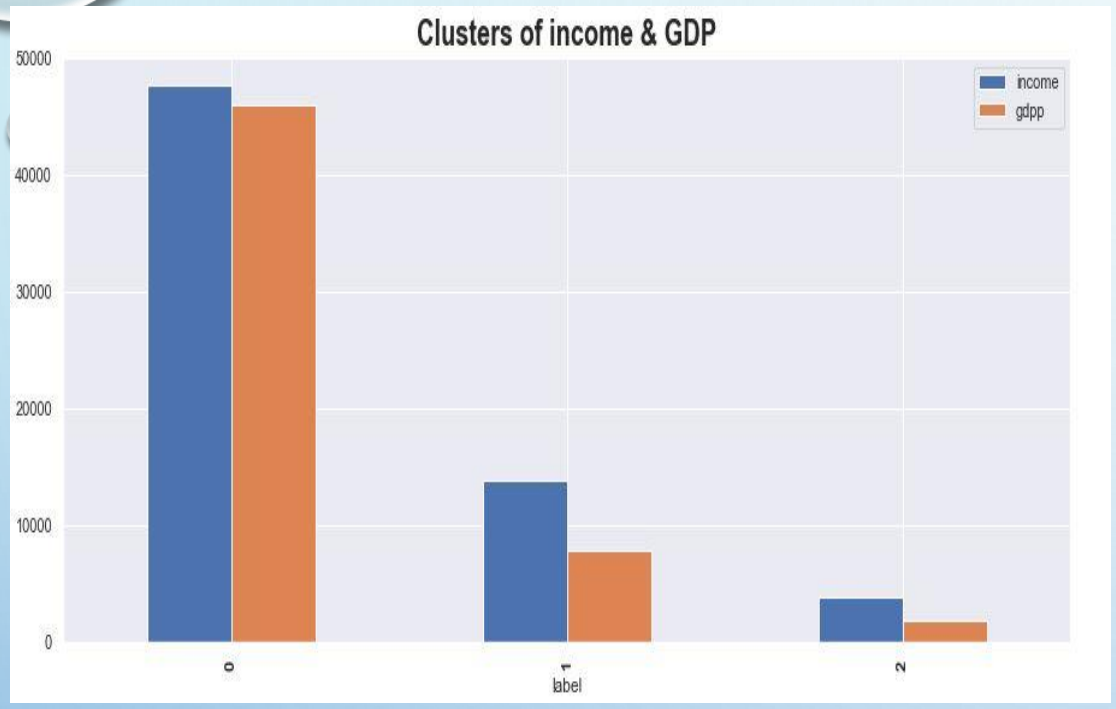
K Means Clustering continued…

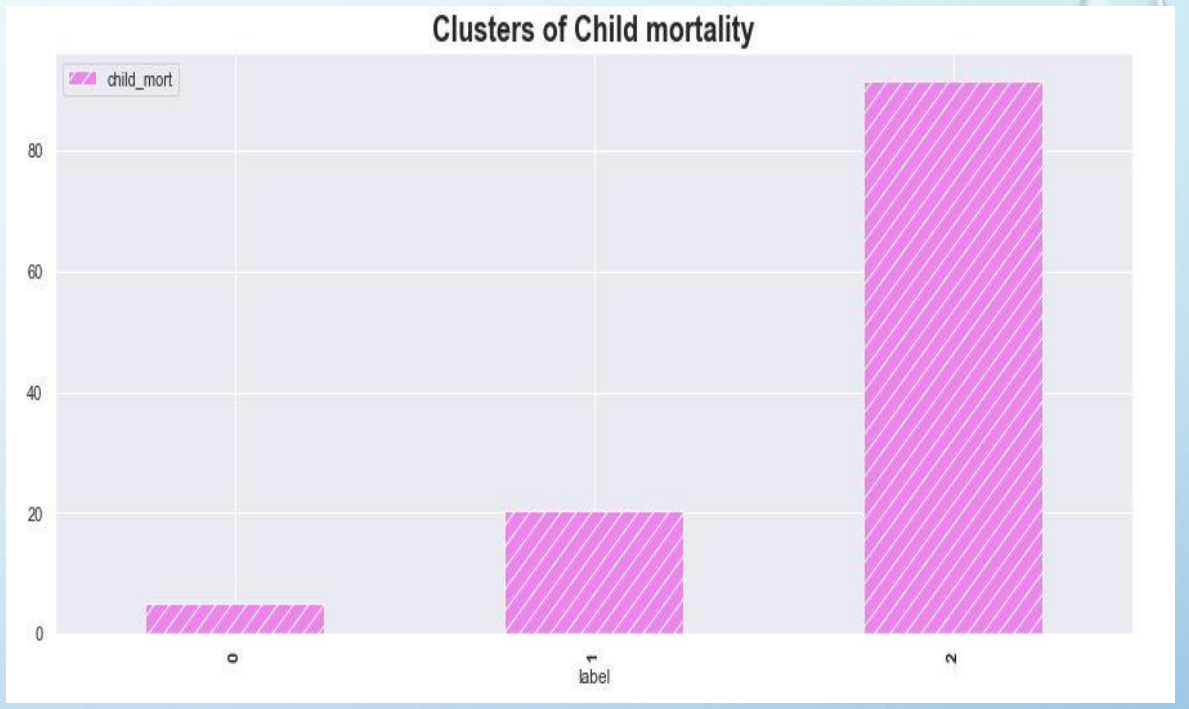
Plots of 3 clusters obtained:



From the three plots we can see that,
Cluster 1 is having high number of data points compared to other two clusters

CLUSTER PROFILING





From cluster profiling in K- means clustering we can see that :

- **Cluster 0 is having the High income, High GDP and very Low child mortality**
- **Cluster 2 is having very Low income, very Low GDP but High child mortality**
- **Cluster 1 is having low income, GDP and less child mortality**

So we can say that countries under cluster 2 are in need of aid.

K MEANS CLUSTERING

Top 10 Countries In Direst Need Of Aid

country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	label
Haiti	208.0000	101.2860	45.7442	428.3140	1500.0000	5.4500	32.1000	3.3300	662.0000	2
Sierra Leone	160.0000	67.0320	52.2690	137.6550	1220.0000	17.2000	55.0000	5.2000	399.0000	2
Chad	150.0000	330.0960	40.6341	390.1950	1930.0000	6.3900	56.5000	6.5636	897.0000	2
Central African Republic	149.0000	52.6280	17.7508	118.1900	888.0000	2.0100	47.5000	5.2100	446.0000	2
Mali	137.0000	161.4240	35.2584	248.5080	1870.0000	4.3700	59.5000	6.5500	708.0000	2
Nigeria	130.0000	589.4900	118.1310	405.4200	5150.0000	41.4780	60.5000	5.8400	2330.0000	2
Niger	123.0000	77.2560	17.9568	170.8680	814.0000	2.5500	58.8000	6.5636	348.0000	2
Angola	119.0000	2199.1900	100.6050	1514.3700	5900.0000	22.4000	60.1000	6.1600	3530.0000	2
Congo, Dem. Rep.	116.0000	137.2740	26.4194	165.6640	609.0000	20.8000	57.5000	6.5400	334.0000	2
Burkina Faso	116.0000	110.4000	38.7550	170.2000	1430.0000	6.8100	57.9000	5.8700	575.0000	2

HIERARCHICAL CLUSTERING

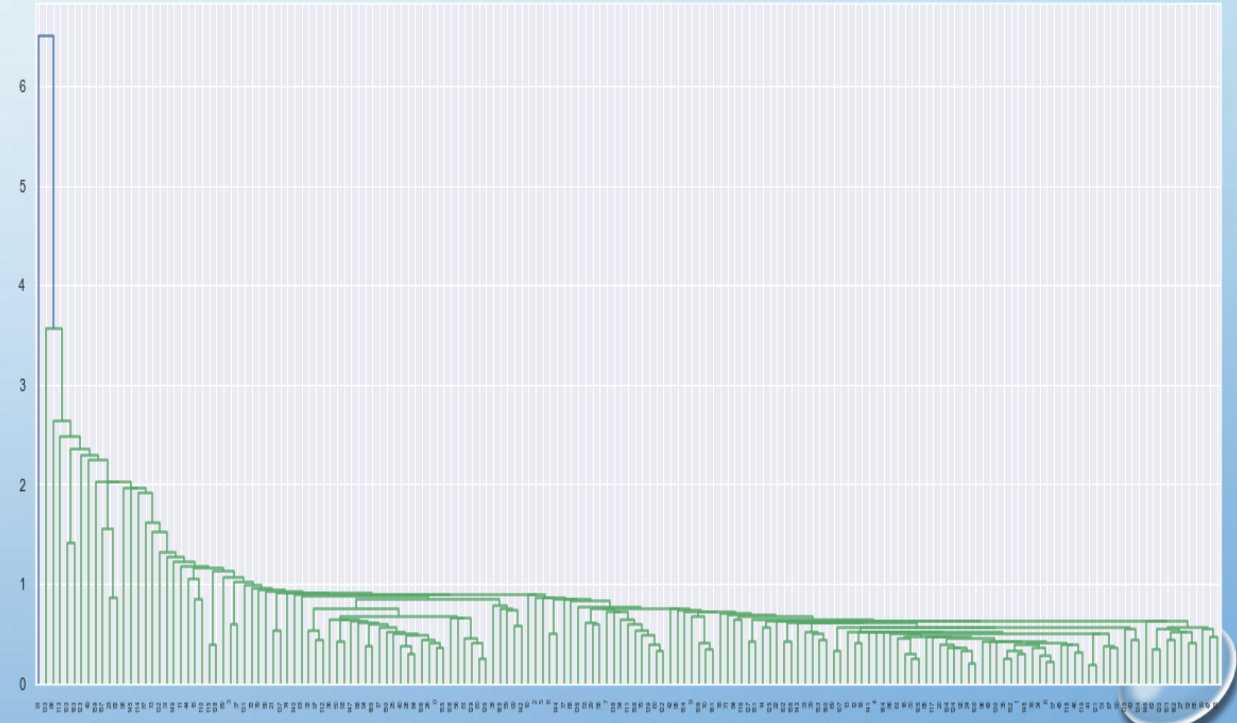
- Hierarchical clustering involves creating clusters that have a predetermined ordering from top to bottom.
- There are two types of hierarchical clustering,

  1.Divisive

  2.Agglomerative.

Single Linkage:

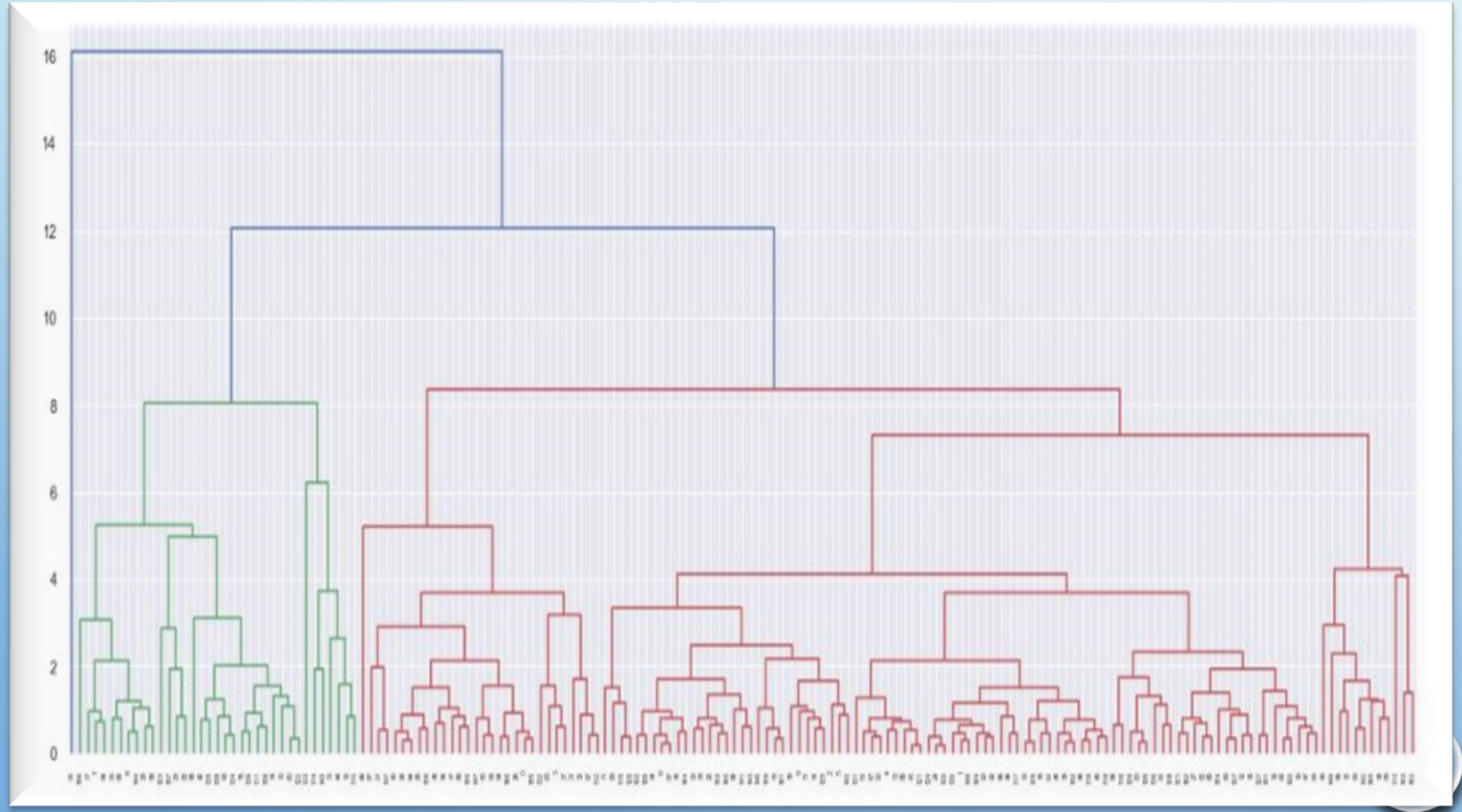
- In single linkage of hierarchical clustering, the distance between two clusters is defined as the shortest distance between two points in each cluster.
- We clearly see, single linkage doesn't produce a good enough result for us to analyse the clusters.
- Hence, we need to go ahead and utilise the complete linkage method and then analyse the clusters once again.



Hierarchical clustering continued…

Complete Linkage

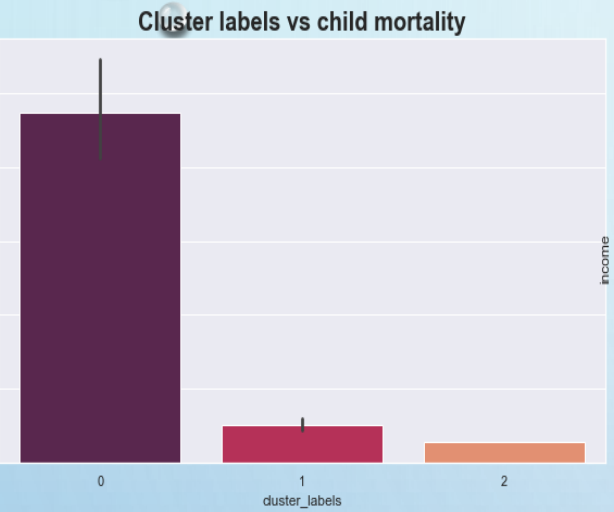
- In complete linkage hierarchical clustering, the distance between two clusters is defined as the longest distance between two points in each cluster.

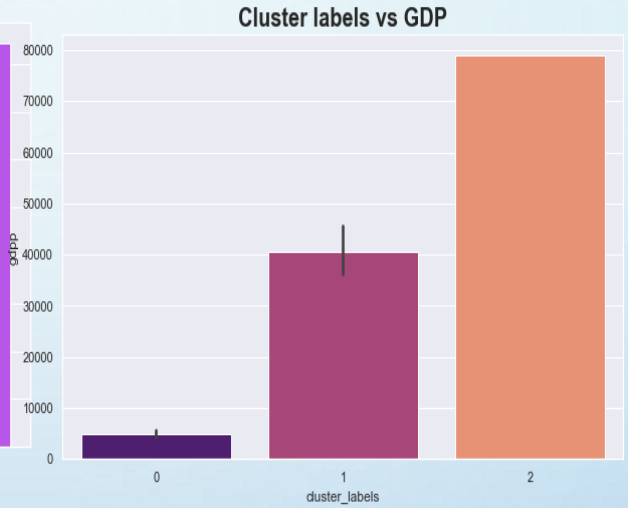


- Now we got the clear dendrogram and its easier to analyse the clusters.
- Lets consider a threshold value of 10. Draw the horizontal line at that height. It cuts 3 vertical lines, all of which represent a cluster.
- So we have 3 clusters now

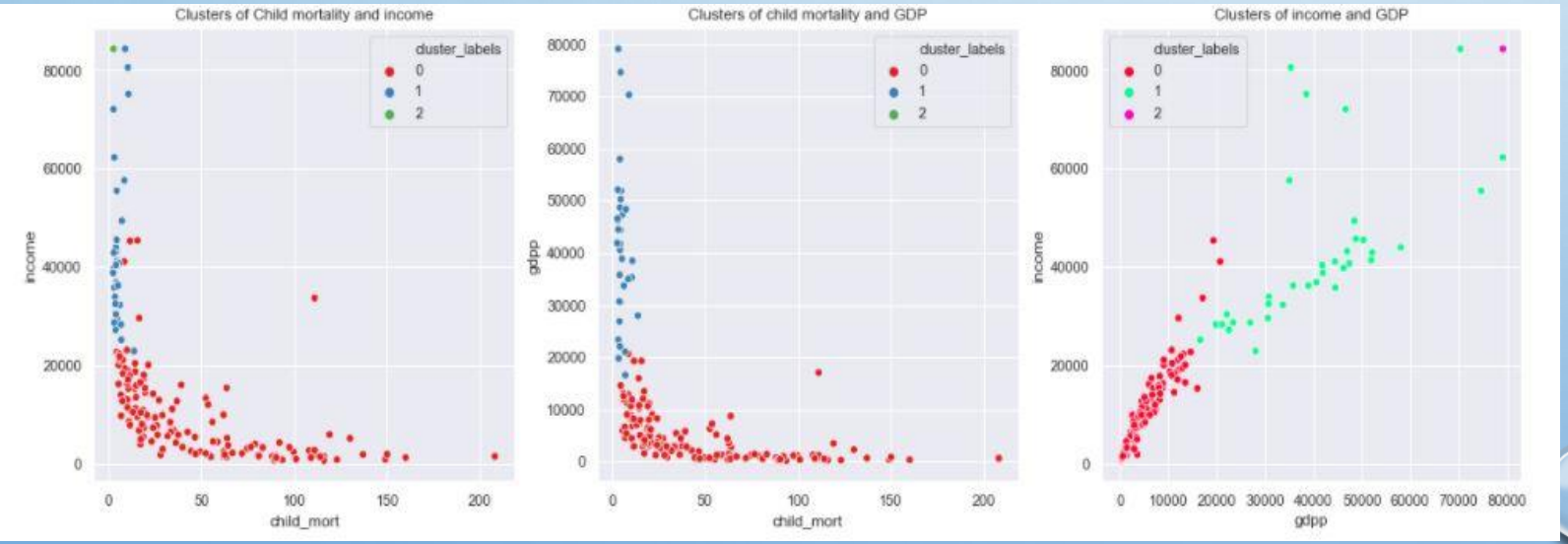
Hierarchical clustering continued…

Plots of 3 clusters obtained:





We can see that the clusters formed are not great. Cluster 2 is having only one data point



From cluster profiling using hierarchical clustering we can see that :

- **Cluster 0 is having the High child mortality, low GDP and very Low child mortality**
- **Cluster 1 is having Low child mortality, moderate income and GDP**
- **Cluster 2 is having very low child mortality, high income and GDP**

HIERARCHICAL CLUSTERING

Top 10 Countries In Direst Need Of Aid

country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	cluster_labels
Haiti	208.0000	101.2860	45.7442	428.3140	1500.0000	5.4500	32.1000	3.3300	662.0000	0
Sierra Leone	160.0000	67.0320	52.2690	137.6550	1220.0000	17.2000	55.0000	5.2000	399.0000	0
Chad	150.0000	330.0960	40.6341	390.1950	1930.0000	6.3900	56.5000	6.5636	897.0000	0
Central African Republic	149.0000	52.6280	17.7508	118.1900	888.0000	2.0100	47.5000	5.2100	446.0000	0
Mali	137.0000	161.4240	35.2584	248.5080	1870.0000	4.3700	59.5000	6.5500	708.0000	0
Nigeria	130.0000	589.4900	118.1310	405.4200	5150.0000	41.4780	60.5000	5.8400	2330.0000	0
Niger	123.0000	77.2560	17.9568	170.8680	814.0000	2.5500	58.8000	6.5636	348.0000	0
Angola	119.0000	2199.1900	100.6050	1514.3700	5900.0000	22.4000	60.1000	6.1600	3530.0000	0
Congo, Dem. Rep.	116.0000	137.2740	26.4194	165.6640	609.0000	20.8000	57.5000	6.5400	334.0000	0
Burkina Faso	116.0000	110.4000	38.7550	170.2000	1430.0000	6.8100	57.9000	5.8700	575.0000	0

We have analysed both K-means and Hierarchical clustering and found clusters formed in both are not identical. The clusters formed in Hierarchical clustering are not great. So, we will proceed with the clusters formed by K-means and based on the information provided by the final clusters we will deduce the final list of countries which are in need of aid.

COUNTRIES BASED ON SOCIO ECONOMIC AND HEALTH FACTORS

- **For the cluster formed in K Mean clustering which is cluster 2, Lets consider the countries in which child mortality is more than average mean value of child mortality in the particular cluster, and income less than the average income, GDP less than average GDP in particular cluster.**

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	label
66	Haiti	208.0000	101.2860	45.7442	428.3140	1500.0000	5.4500	32.1000	3.3300	662.0000	2
132	Sierra Leone	160.0000	67.0320	52.2690	137.6550	1220.0000	17.2000	55.0000	5.2000	399.0000	2
31	Central African Republic	149.0000	52.6280	17.7508	118.1900	888.0000	2.0100	47.5000	5.2100	446.0000	2
112	Niger	123.0000	77.2560	17.9568	170.8680	814.0000	2.5500	58.8000	6.5636	348.0000	2
37	Congo, Dem. Rep.	116.0000	137.2740	26.4194	165.6640	609.0000	20.8000	57.5000	6.5400	334.0000	2
25	Burkina Faso	116.0000	110.4000	38.7550	170.2000	1430.0000	6.8100	57.9000	5.8700	575.0000	2
64	Guinea-Bissau	114.0000	81.5030	46.4950	192.5440	1390.0000	2.9700	55.6000	5.0500	547.0000	2
63	Guinea	109.0000	196.3440	31.9464	279.9360	1190.0000	16.1000	58.0000	5.3400	648.0000	2
106	Mozambique	101.0000	131.9850	21.8299	193.5780	918.0000	7.6400	54.5000	5.5600	419.0000	2
26	Burundi	93.6000	20.6052	26.7960	90.5520	764.0000	12.3000	57.7000	6.2600	231.0000	2

INFERENCES

- From the EDA performed we could see that income, GDP and child mortality are the major three variables need to be focused
- In K means clustering we got cluster 2 is having very low income, very low GDP but high child mortality. So we concluded that countries under cluster 2 are in need of aid.
- In hierarchical clustering we saw that cluster 0 is having the high child mortality, low Gdp and very low child mortality.
- The clusters formed in hierarchical clustering were not that good. So we went on to consider cluster formed in K means clustering. And got top countries with high child mortality, low GDP and low income

LIST OF COUNTRIES TO BE FOCUSED ON

From 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 Republic**
10. **Burkina Faso**

From Socio Economic & Health Factors:

1. **Haiti**
2. **Sierra Leone**
3. **Central African Republic**
4. **Niger**
5. **Central Dem, Republic**
6. **Burkina Faso**
7. **Guinea - Bissau**
8. **Guinea**
9. **Mozambique**
10. **Burundi**

RECOMMENDATIONS

- **HELP International CEO must focus on the countries where the people have less income.**
- **Less income in turn affects the total GDP per capita of that country. So focus more on Low GDP countries.**
- **Countries with Low GDP can focus more on imports and exports. And reduce total fertility. That might help a little economically.**
- **Import and Export of goods & services have high influence of GDP of the country.**
- **Focus more on countries where the child mortality is high**
- **There are some countries which spend well on health for the people living in that country. For ex: US. Such countries can be skipped. And focus More On Burundi, Congo, Dem. Rep where the total health spending is too less.**
- **If the total fertility is less the life expectancy is more. Haiti is the country having very low life expectancy, and high child mortality. Its good to have less children per woman, so that they could be looked after well.**