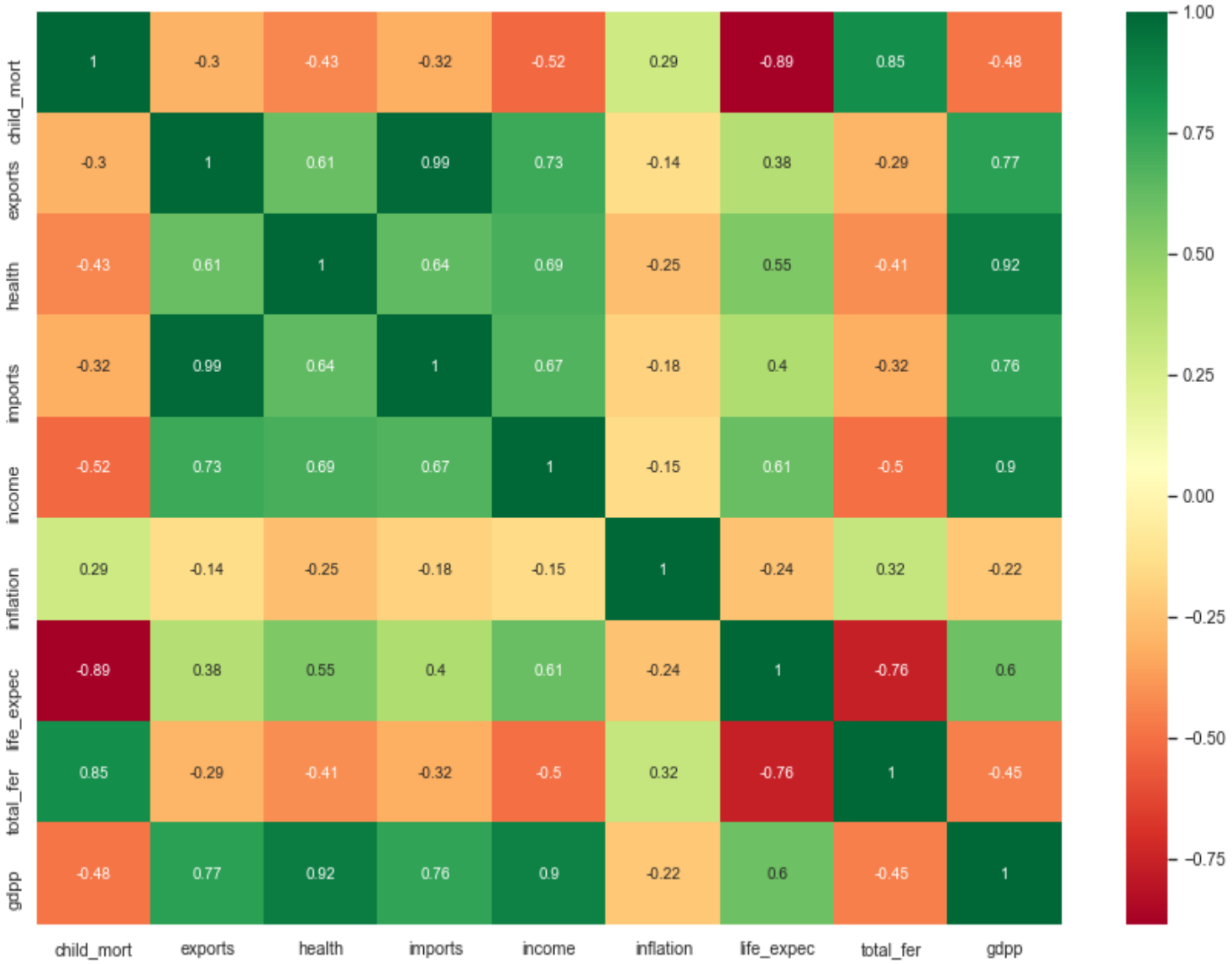
Clustering Case Study

- 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 programs, 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.
- The objective is to categories the countries using some socio-economic and health factors that determine the overall development of the country and suggest to CEO the list of country to focus on.

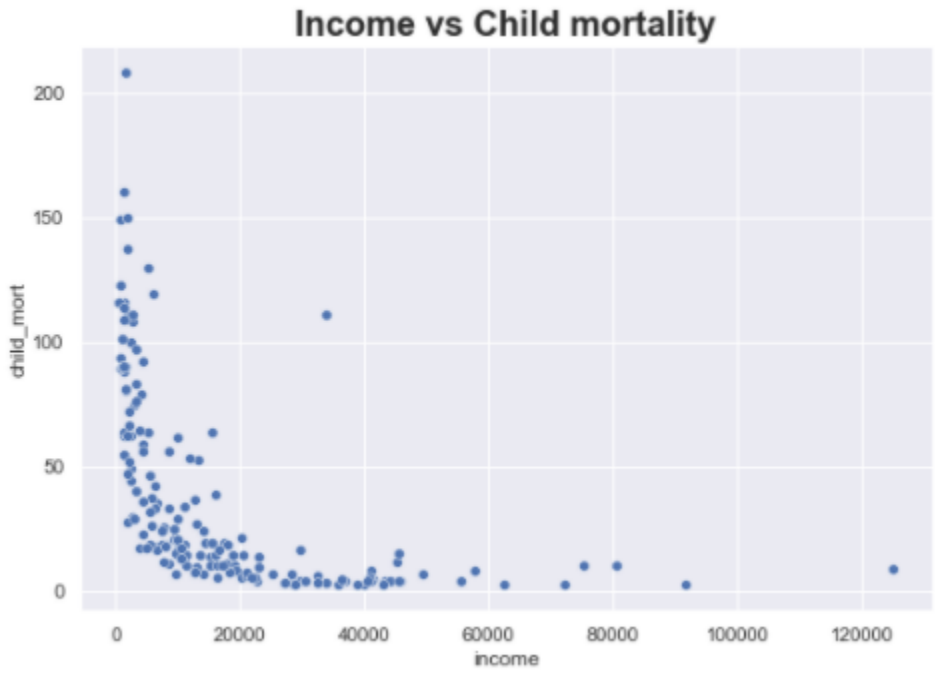
- Created by-
  - Sandip Patra

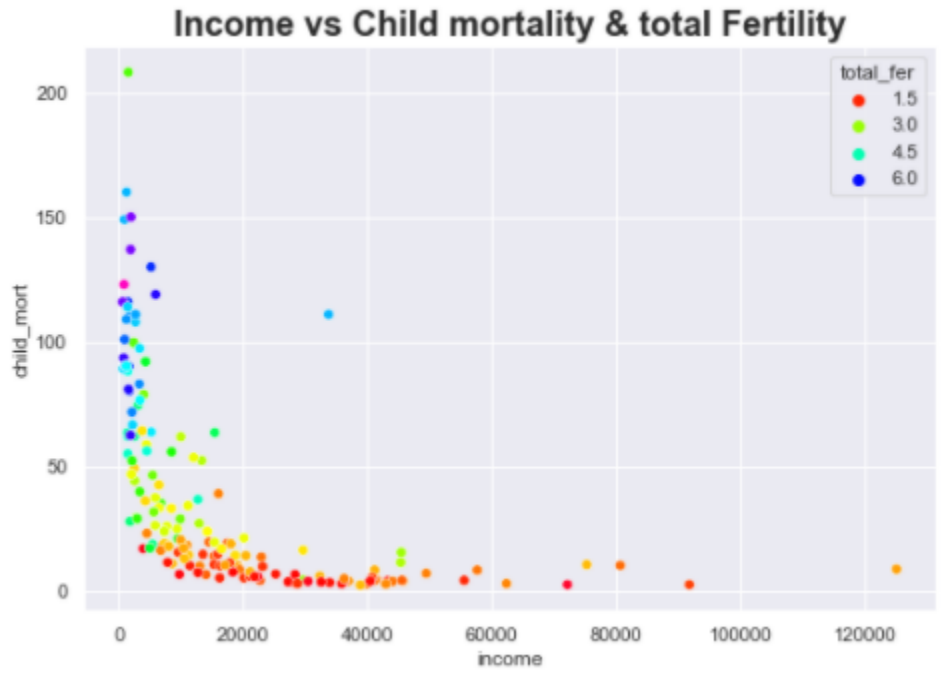
Correlation Analysis



- Points to be concluded:
- Child mortality is also in strong negative correlation with life expectancy.
- Child mortality is highly correlated with total fertility with the correlation coefficient of 0.85
- Life expectancy has strong negative correlation with total fertility of about -0.76

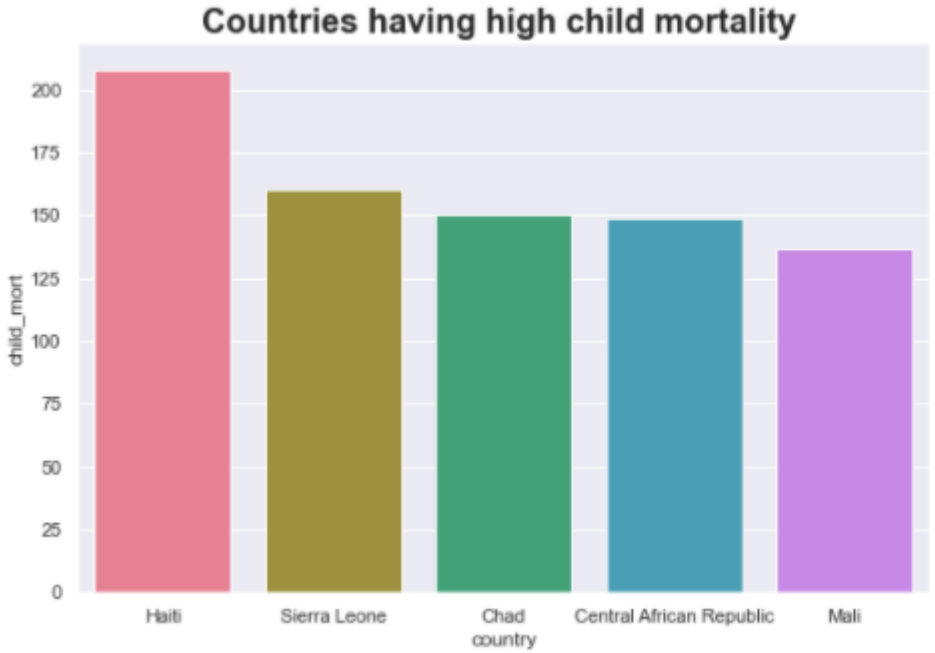
Income vs Child Mortality Analysis

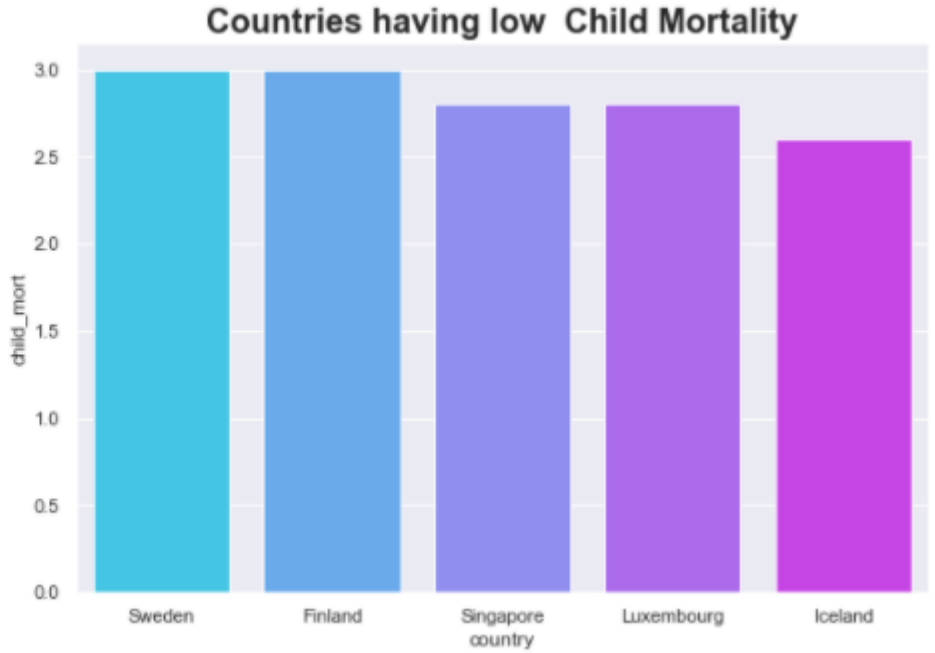




- From the plots above We can see that low income people have high child mortality, which means death of children under age 5 is more, where there is a low income
- In the second plot we can see that, high fertility rate for a woman and low income have high child mortality

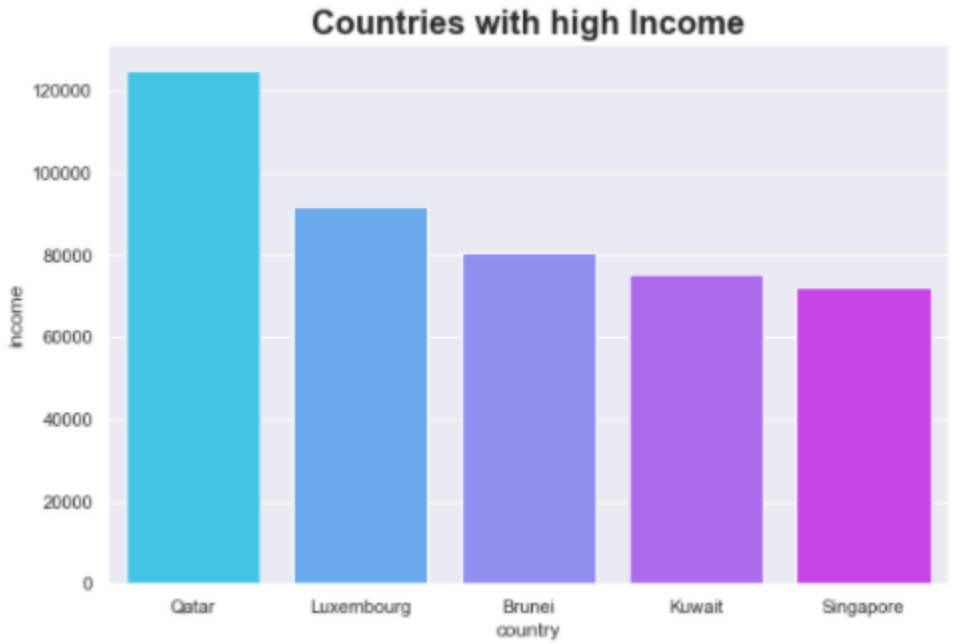
Country vs Child Mortality

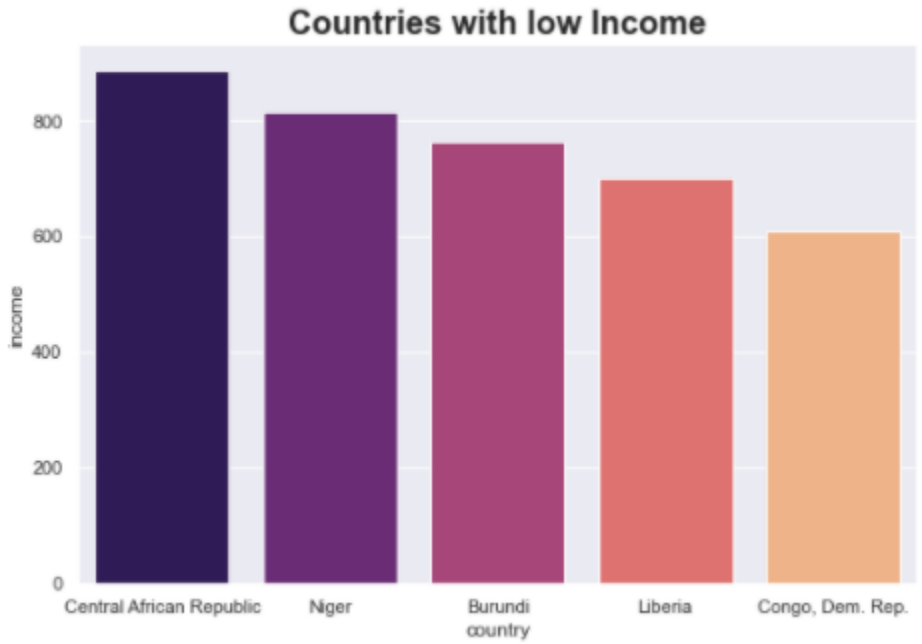




- Haiti is a country with highest child mortality of around 208 out of 1000 live births. Next comes Sierra Leone
- Iceland, Luxembourg, Singapore are the countries with less child mortality

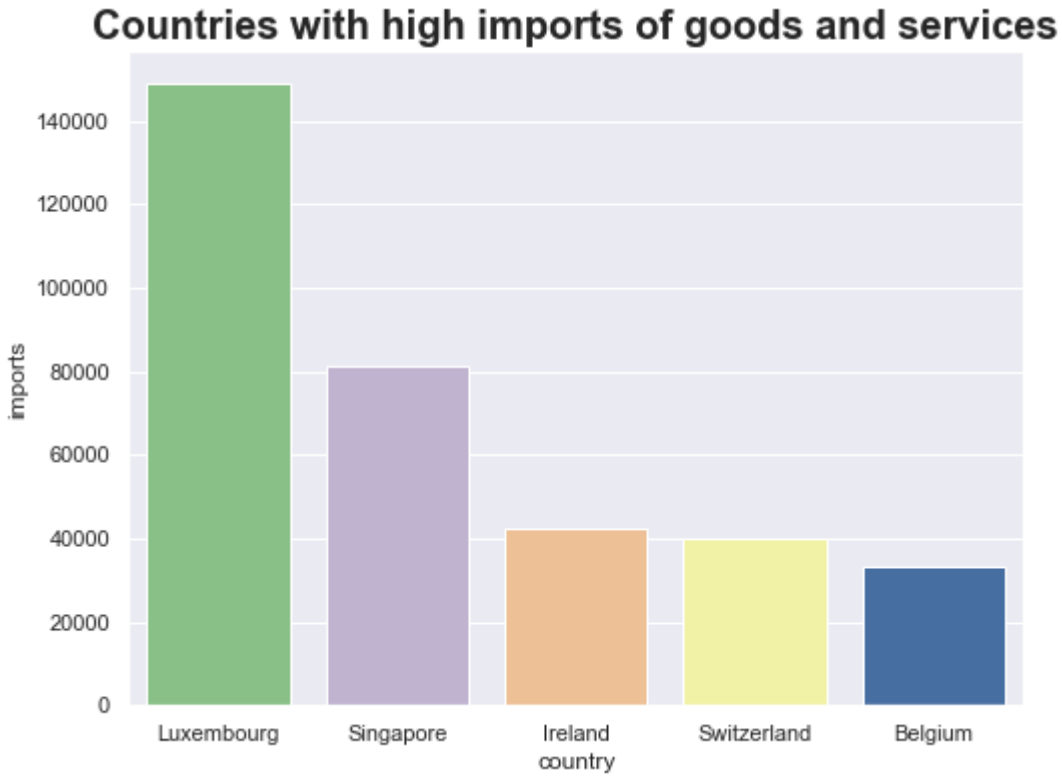
Country vs Income





- Net income per person is more in Qatar which is 125000
- Congo, Democratic Republic and Liberia has less income.

Country vs Imports



- 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

K-mean clustering

```
df_kmean = pd.concat([df_kmean, label], axis =1)
df_kmean.head()
```

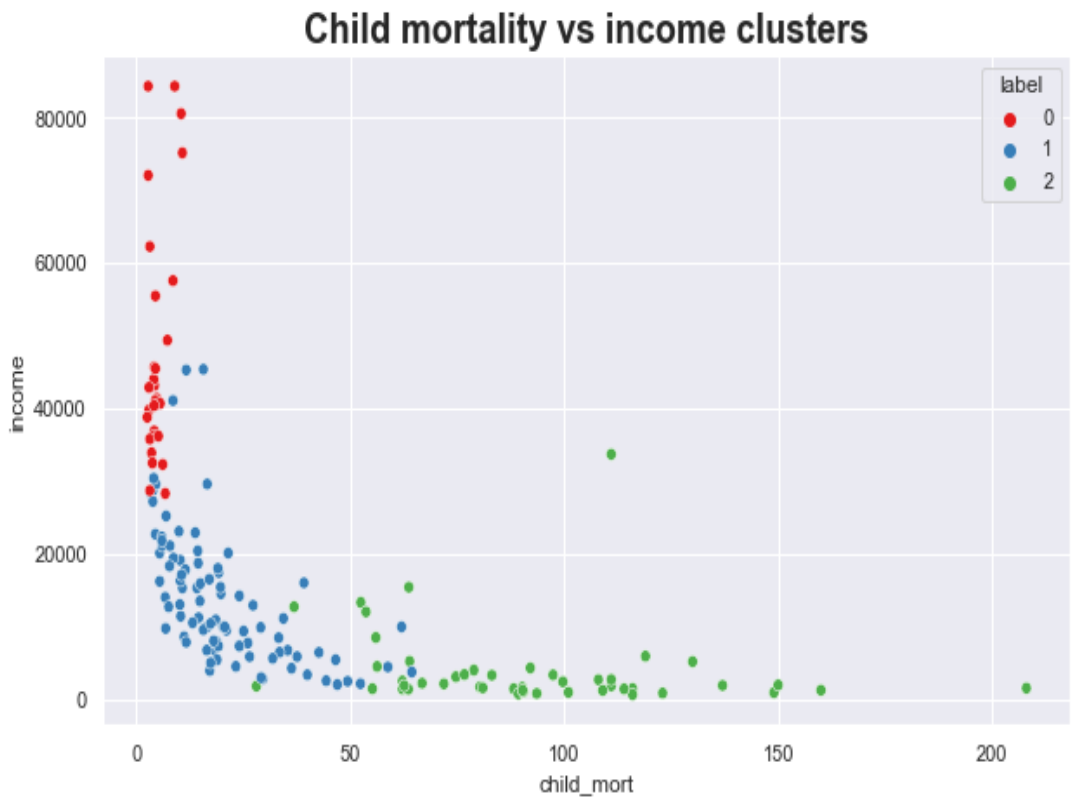
	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	label
0	Afghanistan	90.2000	55.3000	41.9174	248.2970	1610.0000	9.4400	56.2000	5.8200	553.0000	2
1	Albania	16.6000	1145.2000	267.8950	1987.7400	9930.0000	4.4900	76.3000	1.6500	4090.0000	1
2	Algeria	27.3000	1712.6400	185.9820	1400.4400	12900.0000	16.1000	76.5000	2.8900	4460.0000	1
3	Angola	119.0000	2199.1900	100.6050	1514.3700	5900.0000	22.4000	60.1000	6.1600	3530.0000	2
4	Antigua and Barbuda	10.3000	5551.0000	735.6600	7185.8000	19100.0000	1.4400	76.8000	2.1300	12200.0000	1

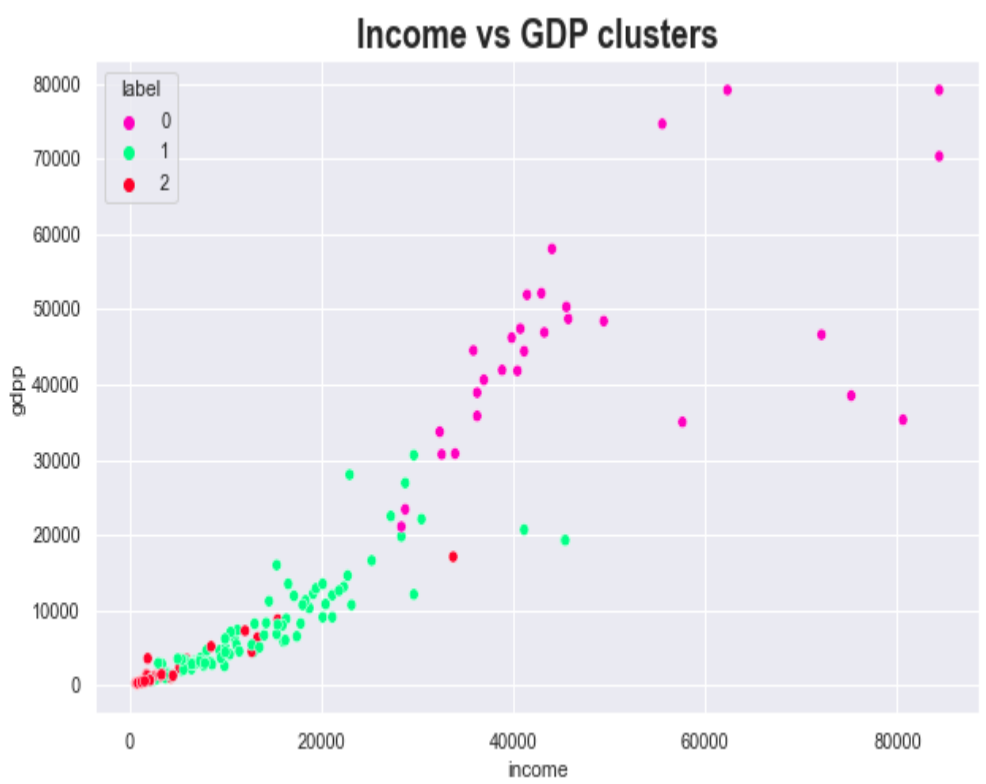
```
df_kmean.label.value_counts()
```

```
1    90
2    48
0    29
Name: label, dtype: int64
```

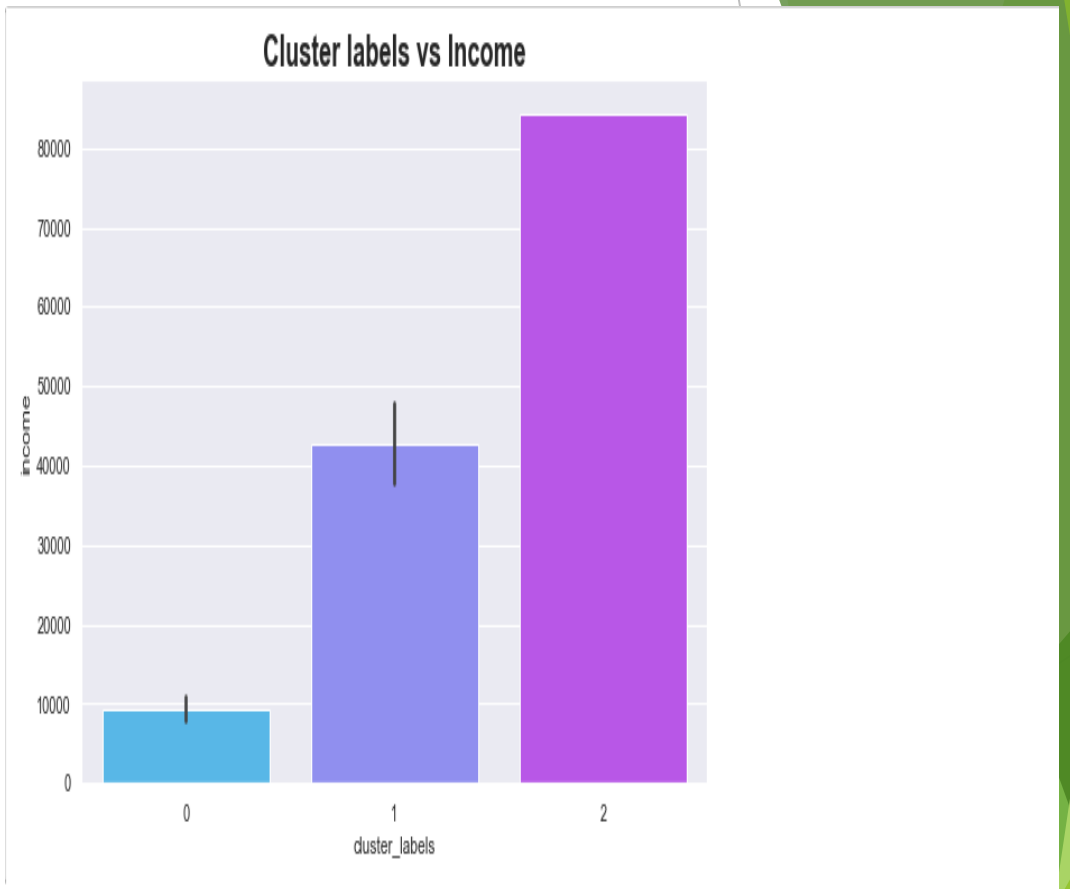
- **Cluster 1 is having high number of data points compared to other two clusters**

Plotting the clusters





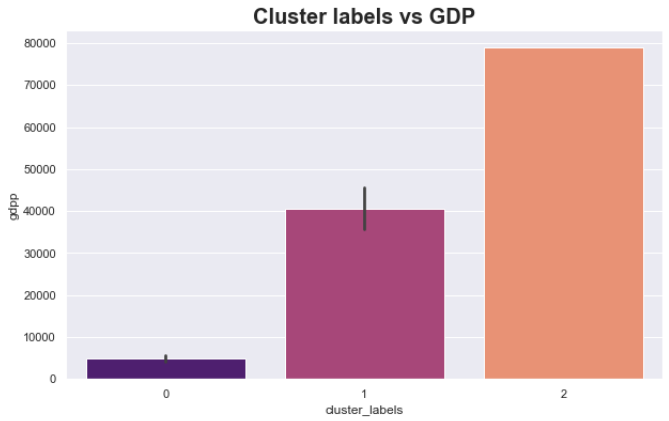
Hierarchical Clustering



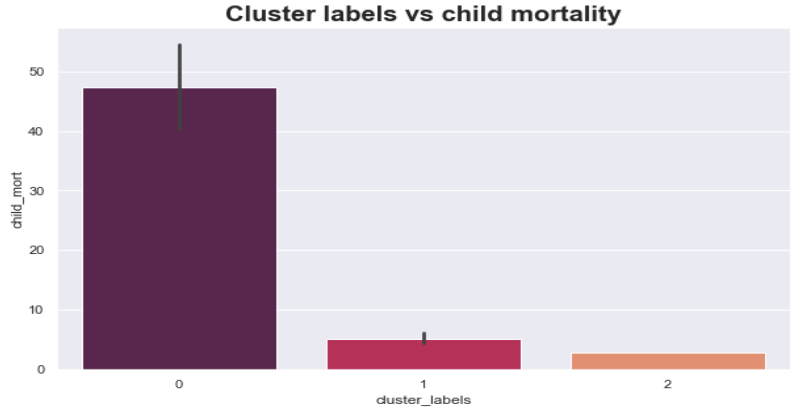
Cluster 0 is having high child mortality

Cluster 1 and 2 have very low child mortality

Cluster 2 is having more income and cluster 1 has modertate income, cluster 0 has very least income

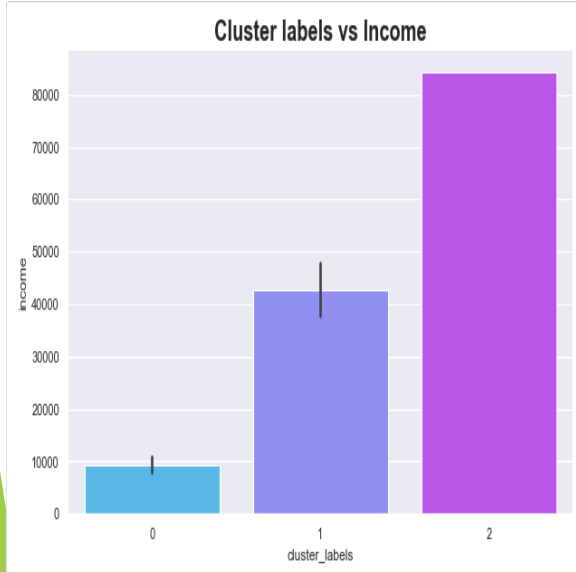


Cluster 0 have very low GDP, cluster 2 have high GDP and cluster 1 has moderte GDP



Cluster 0 is having high child mortality

Cluster 1 and 2 have very low child mortality



Cluster 2 is having more income and cluster 1 has modertate income, cluster 0 has very least income

Observations-

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

From K means clustering we got better clusters compared to Hierarchical clustering.

Cluster 2 is the better cluster we got with High child mortality, low income and low GDP

Final list of country we got are:

```
Kmean.head()
```

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	label
0	Afghanistan	90.2000	55.3000	41.9174	248.2970	1610.0000	9.4400	56.2000	5.8200	553.0000	2
3	Angola	119.0000	2199.1900	100.6050	1514.3700	5900.0000	22.4000	60.1000	6.1600	3530.0000	2
17	Benin	111.0000	180.4040	31.0780	281.9760	1820.0000	0.8850	61.8000	5.3600	758.0000	2
21	Botswana	52.5000	2768.6000	527.0500	3257.5500	13300.0000	8.9200	57.1000	2.8800	6350.0000	2
25	Burkina Faso	116.0000	110.4000	38.7550	170.2000	1430.0000	6.8100	57.9000	5.8700	575.0000	2

```
Kmean=Kmean.sort_values(by=['child_mort','income','gdpp'], ascending=[False,True,True])
Kmean.head()
```

	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
32	Chad	150.0000	330.0960	40.6341	390.1950	1930.0000	6.3900	56.5000	6.5636	897.0000	2
31	Central African Republic	149.0000	52.6280	17.7508	118.1900	888.0000	2.0100	47.5000	5.2100	446.0000	2
97	Mali	137.0000	161.4240	35.2584	248.5080	1870.0000	4.3700	59.5000	6.5500	708.0000	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 five countries with High child mortality,Low GDP and Low income
- Then we looked for the countries based on socio economic factors

```
print('Top 5 Countries based on K means clustering:')
Kmean[['country']].head()
```

```
Top 5 Countries based on K means clustering:
```

	country
66	Haiti
132	Sierra Leone
32	Chad
31	Central African Republic
97	Mali

```
print('Countries based on socio economic and health factors:')
A_countries
```

```
Countries based on socio economic and health factors:

0    Haiti
1    Sierra Leone
2    Central African Republic
3    Niger
4    Congo, Dem. Rep.
5    Burkina Faso
6    Guinea-Bissau
7    Guinea
8    Mozambique
9    Burundi
Name: country, dtype: object
```

Recommendations

- From the examination performed, We can see that low income individuals have high child mortality, so CEO should focus more on low income nations
- We could likewise observe Low GDP per capita nations are not having a lot of import and export of good and services. Those nations additionally should be focused.
- There are a few nations which spend well on wellbeing for individuals living in that nation. For ex: US. Such nations can be skipped. Furthermore, center more around Burundi, Congo, Dem. Rep where the absolute wellbeing spending is excessively less.
- In the event that the all-out fruitfulness is less the existence expectancy is more. Haiti is the nation having extremely low life anticipation, and high child mortality. Its great to have less kids per woman, so that they could be taken care of well.

Thank you