

ASSIGNMENT : HELP International using K-mean clustering

SUBMITTED BY : Nidhi Sharma

Overview

- 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. 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.
- My job is:-
 - To categorise the countries using some socio-economic and health factors that determine the overall development of the country.
 - We need to suggest the countries which the CEO needs to focus on the most.

AIM/TASK: Identify top 10 countries that are direst need of aid.

Job : Categorise the countries using some socio-economic and health factors that determine the overall development of the country.

Suggestions : atleast 5 countries which the CEO needs to focus on the most.

Steps followed are:-

- 1. Read and understand the data
- 2. Clean the data
- 3. Visualization of data
- 3. Prepare the data for modelling
- 4. Hopkins Statistics Test
- 5. Modelling
- 6. Final analysis

1. Read,Data understanding and representation

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

child_mort: Death of children under 5 years of age per 1000 live births

Income: Net income per person

Inflation: The measurement of the annual growth rate of the Total GDP

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.

This fig shows the data has 10 columns and 167 rows

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
0	Afghanistan	90.2	10.0	7.58	44.9	1610	9.44	56.2	5.82	553
1	Albania	16.6	28.0	6.55	48.6	9930	4.49	76.3	1.65	4090
2	Algeria	27.3	38.4	4.17	31.4	12900	16.10	76.5	2.89	4460
3	Angola	119.0	62.3	2.85	42.9	5900	22.40	60.1	6.16	3530
4	Antigua and Barbuda	10.3	45.5	6.03	58.9	19100	1.44	76.8	2.13	12200
...
162	Vanuatu	29.2	46.6	5.25	52.7	2950	2.62	63.0	3.50	2970
163	Venezuela	17.1	28.5	4.91	17.6	16500	45.90	75.4	2.47	13500
164	Vietnam	23.3	72.0	6.84	80.2	4490	12.10	73.1	1.95	1310
165	Yemen	56.3	30.0	5.18	34.4	4480	23.60	67.5	4.67	1310
166	Zambia	83.1	37.0	5.89	30.9	3280	14.00	52.0	5.40	1460

167 rows × 10 columns

Converting the exports, imports and health variables to actual values:-

Converting imports, exports and health spending from percentage values to actual values of their GDP per capita. Because the percentage values don't give a clear picture of that country. For example Austria and Belarus have almost same exports % (Austria=51.3, Belarus= 51.4) but their gdpp has a huge gap (Austria=46900, Belarus= 6030) which doesn't give an accurate idea of which country is more developed than the other.

```
# Converting exports, imports and health spending percentages to absolute values.  
country['exports'] = country['exports'] * country['gdpp']/100  
country['imports'] = country['imports'] * country['gdpp']/100  
country['health'] = country['health'] * country['gdpp']/100  
country
```

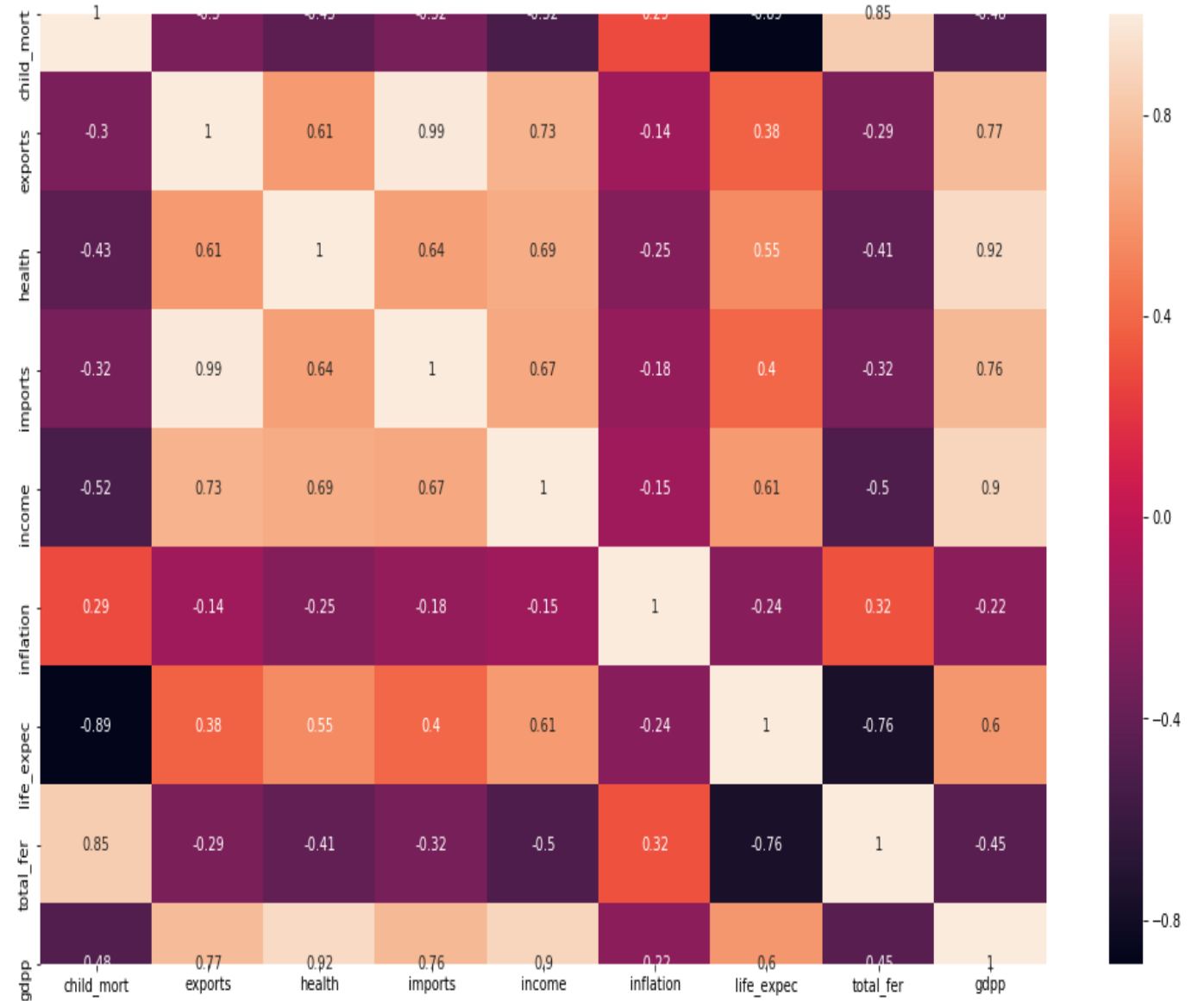
	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
0	Afghanistan	90.2	55.30	41.9174	248.297	1610	9.44	56.2	5.82	553
1	Albania	16.6	1145.20	267.8950	1987.740	9930	4.49	76.3	1.65	4090
2	Algeria	27.3	1712.64	185.9820	1400.440	12900	16.10	76.5	2.89	4460
3	Angola	119.0	2199.19	100.6050	1514.370	5900	22.40	60.1	6.16	3530
4	Antigua and Barbuda	10.3	5551.00	735.6600	7185.800	19100	1.44	76.8	2.13	12200
...
162	Vanuatu	29.2	1384.02	155.9250	1565.190	2950	2.62	63.0	3.50	2970
163	Venezuela	17.1	3847.50	662.8500	2376.000	16500	45.90	75.4	2.47	13500
164	Vietnam	23.3	943.20	89.6040	1050.620	4490	12.10	73.1	1.95	1310
165	Yemen	56.3	393.00	67.8580	450.640	4480	23.60	67.5	4.67	1310
166	Zambia	83.1	540.20	85.9940	451.140	3280	14.00	52.0	5.40	1460

EDA

correlation (Visualization of the data)

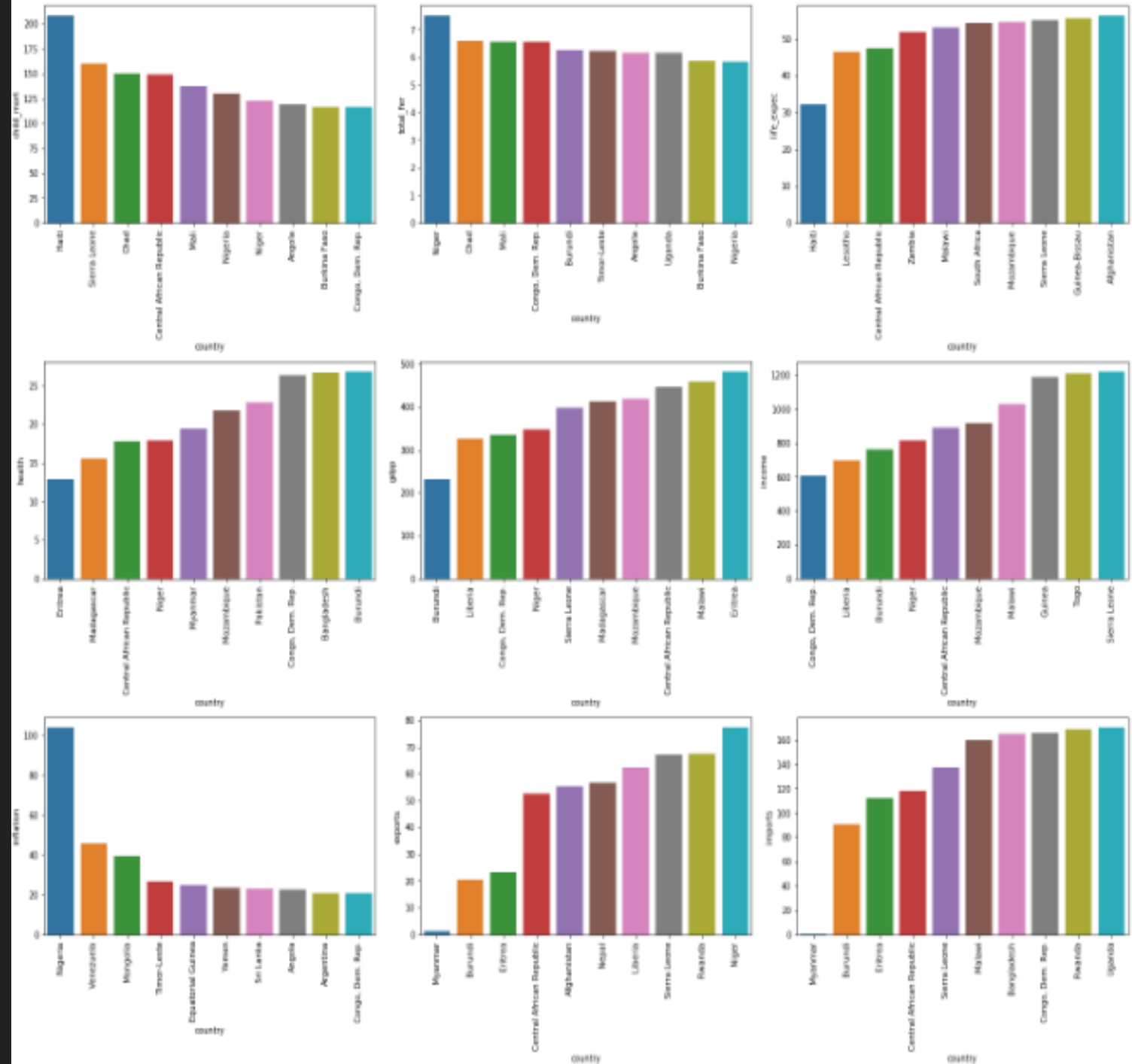
Inference:

- imports and exports are highly correlated with correlation of 0.99
- health and gdpp are highly correlated with correlation of 0.92
- income and gdpp are highly correlated with correlation of 0.9
- child_mortality and life_expencyency are highly correlated with correlation of -0.89
- child_mortality and total_fertility are highly correlated with correlation of 0.85
- gdpp and exports are highly correlated with correlation of 0.77
- gdpp and imports are highly correlated with correlation of 0.76
- life_expencyency and total_fertility are highly correlated with correlation of -0.76
- income and exports are highly correlated with correlation of 0.73



Univariate analysis of the data

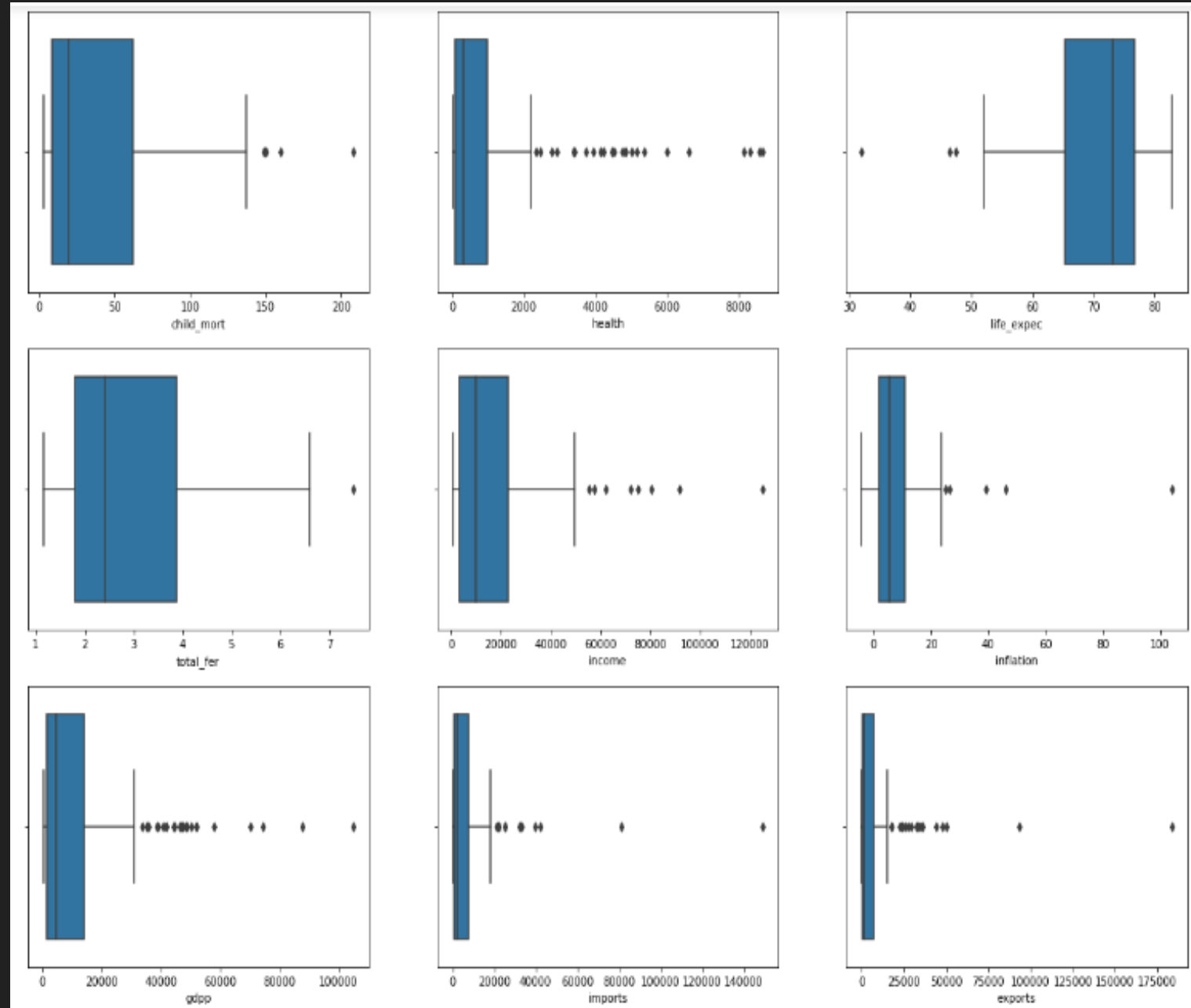
We need to choose the countries that are in the direct need of aid. Hence, we need to identify those countries with using some socio-economic and health factors that determine the overall development of the country. **Top 10 countries based on every column**



Univariate analysis of the data

BOX PLOT

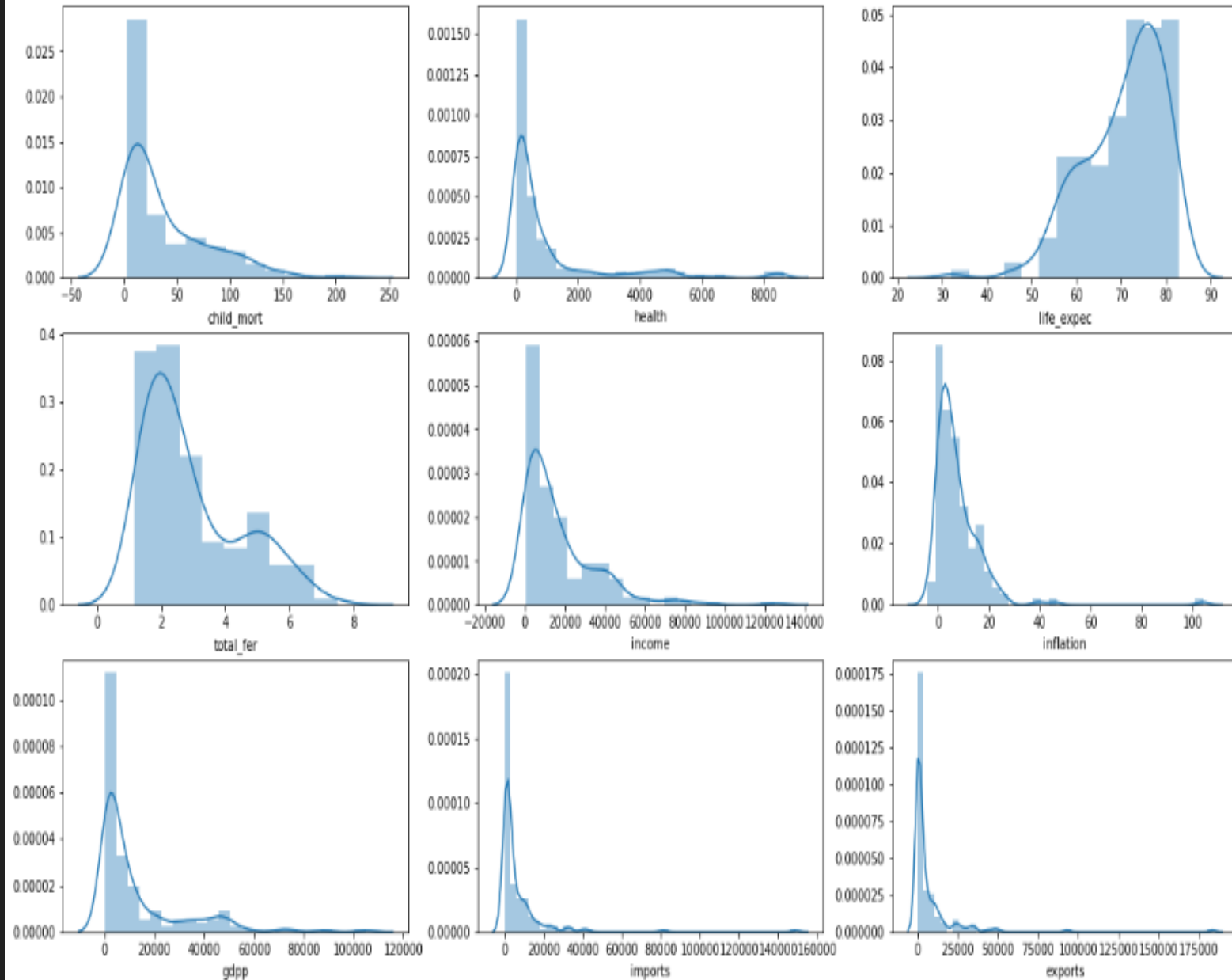
We can see outliers



Univariate analysis of the data

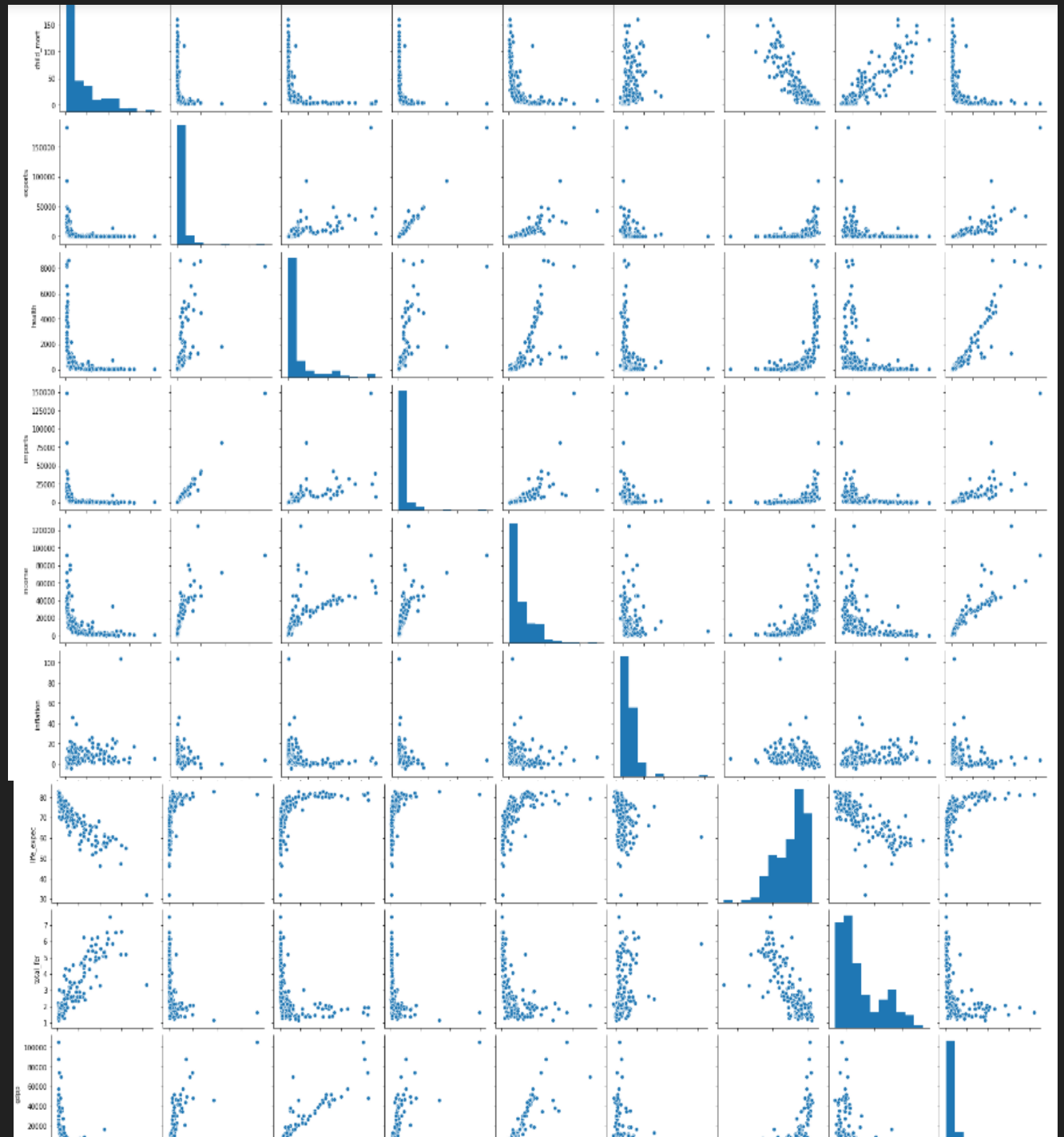
DIST PLOT

We can see
distribution of data



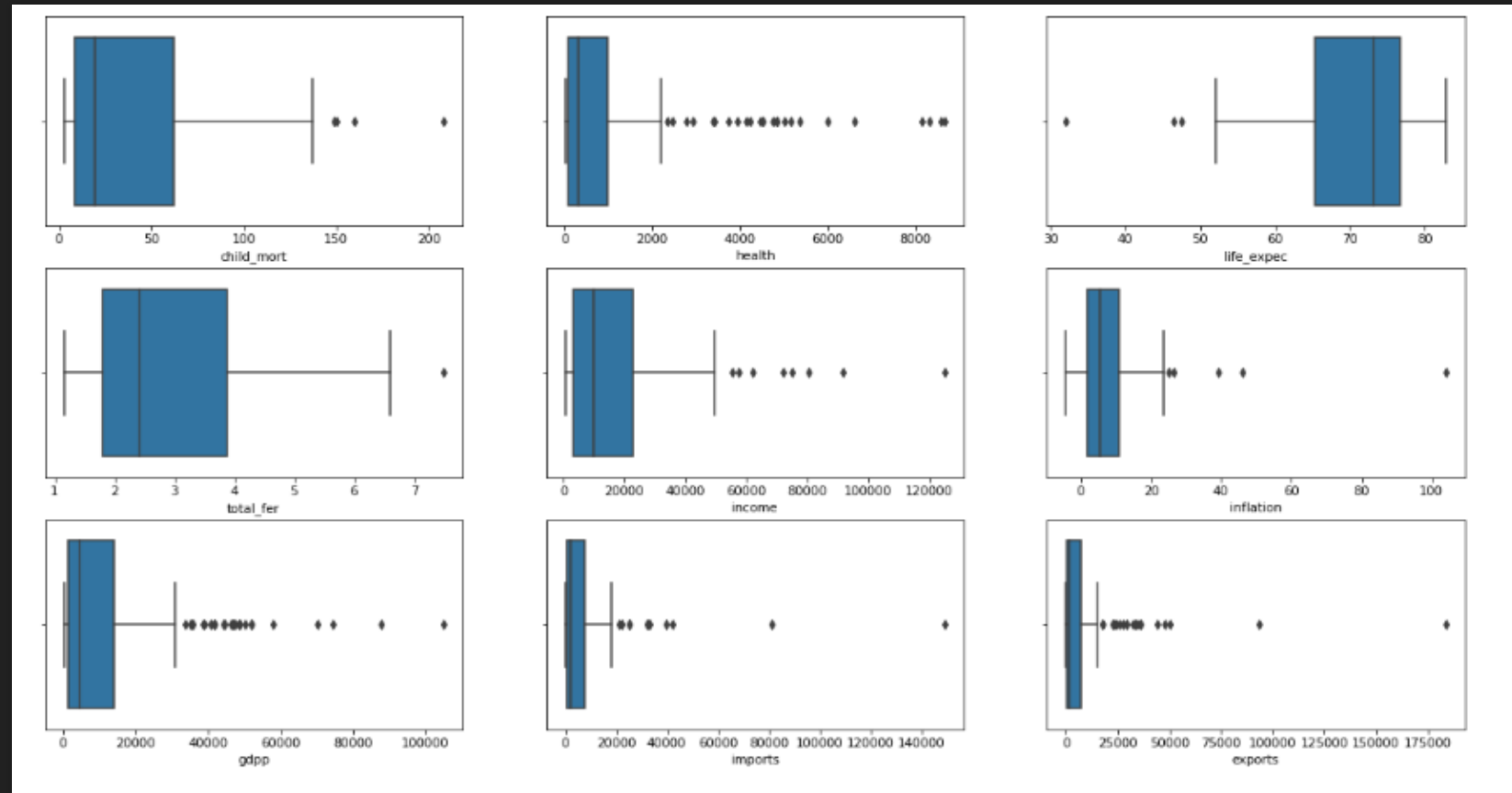
Bivariate analysis of the data

PAIR PLOT



Checking Outliers in the data

➤ **Using capping technique**

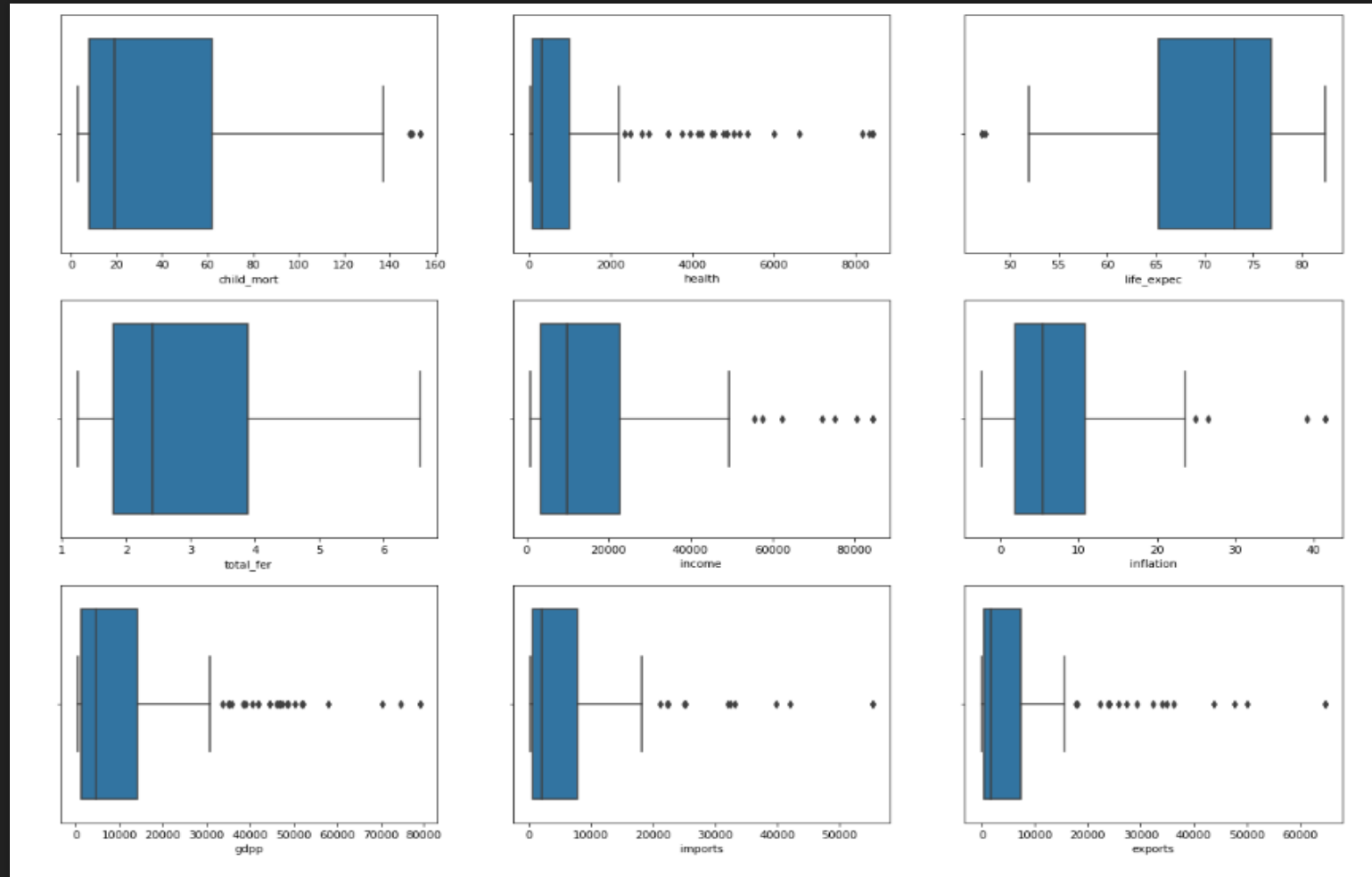


Outlier Treatment

Keeping in mind we need to identify backward countries based on socio economic and health factors. We will cap the outliers to values accordingly for analysis.

Choose capping technique over IQR technique
Assumption : below 1% data will cap in 1%
Above 99% data will cap in 99%

After capping done , few outliers got removed , have to work with these outliers now



Hopkins Check

The Hopkins statistic (introduced by Brian Hopkins and John Gordon Skellam) is a statistic which gives a value which indicates the cluster tendency, in other words: how well the data can be clustered.

If the value is between $\{0.01, \dots, 0.3\}$, the data is regularly spaced.

If the value is around 0.5, it is random.

If the value is between $\{0.7, \dots, 0.99\}$, it has a high tendency to cluster.

Inference: 0.92 is a good Hopkins score for Clustering.

Scaling

The Euclidean distance is calculated by taking the square root of the sum of the squared differences between observations. This distance can be greatly affected by differences in scale among the variables.

So I performed scaling of the variables.

	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
0	1.344012	-0.569638	-0.565164	-0.598844	-0.851772	0.263649	-1.693799	1.926928	-0.702314
1	-0.547543	-0.473873	-0.439335	-0.413679	-0.387025	-0.375251	0.663053	-0.865911	-0.498775
2	-0.272548	-0.424015	-0.484946	-0.476198	-0.221124	1.123260	0.686504	-0.035427	-0.477483
3	2.084186	-0.381264	-0.532486	-0.464070	-0.612136	1.936405	-1.236499	2.154642	-0.531000
4	-0.709457	-0.086754	-0.178874	0.139659	0.125202	-0.768917	0.721681	-0.544433	-0.032079

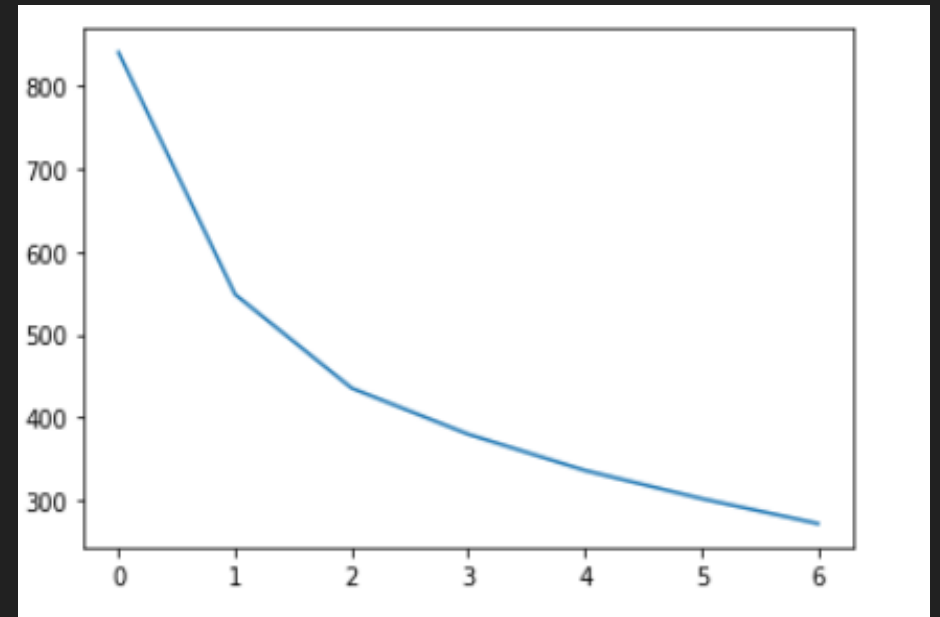
Finding the Optimal Number of Clusters

Method 1: Elbow-Curve/SSD to get the right no. of clusters

- A fundamental step for any unsupervised algorithm is to determine the optimal number of clusters into which the data may be clustered. The Elbow Method is one of the most popular methods to determine this optimal value of k .

Conclusion:

After observing this curve I selected my optimum value of cluster as 3.



Finding the Optimal Number of Clusters

Method 2: Silhouette Analysis

Before proceeding into clustering the data, we find out the optimal number of clusters by the following two methods

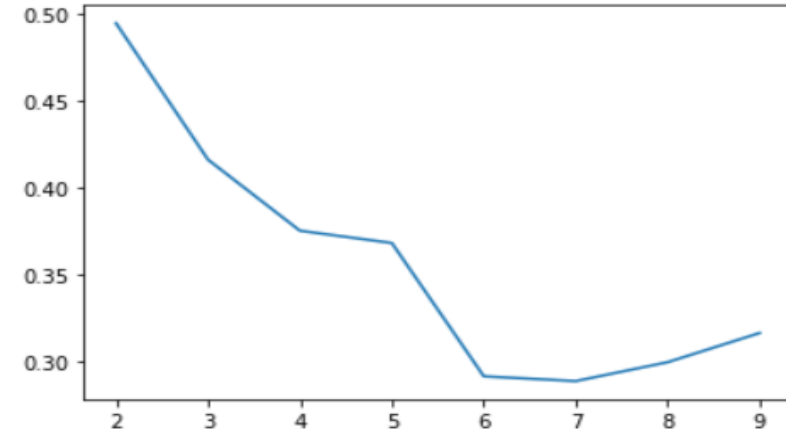
- The value of the silhouette score range lies between -1 to 1.
- A score closer to 1 indicates that the data point is very similar to other data points in the cluster,
- A score closer to -1 indicates that the data point is not similar to the data points in its cluster.

Conclusion:

The silhouette score reaches a peak at around 3 clusters indicating that it might be the ideal number of clusters. (k=3)

Also I observed this curve and silhouette score and then I selected my optimum value of cluster as 3.

```
For n_clusters=2, the silhouette score is 0.49455941306658363
For n_clusters=3, the silhouette score is 0.4160717022232288
For n_clusters=4, the silhouette score is 0.3751726233026849
For n_clusters=5, the silhouette score is 0.35136139823468543
For n_clusters=6, the silhouette score is 0.298155556685058
For n_clusters=7, the silhouette score is 0.3070589058368767
For n_clusters=8, the silhouette score is 0.2812050335739004
```



Cluster profiling:-

- From the business understanding we have learnt that Child_Mortality, Income, Gdpp are some important factors which decides the development of any country. Hence, we will proceed with cluster profiling by using these 3 variables.
- - gdpp: (The GDP per capita) Calculated as the Total GDP divided by the total population.
- - child_mort: Death of children under 5 years of age per 1000 live births.
- - income: Net income per person. g technique

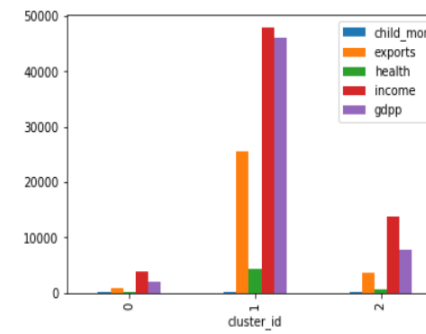
K-Mean clustering

Clusters made using cluster_ids variable

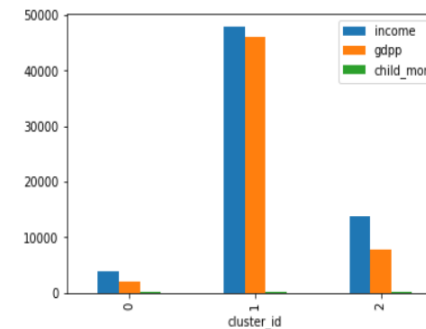
By using K= 3, Analysis of clusters formed from k mean on basis of 3 columns i.e income, gdpp and child_mort

Inference:- Income and Gdpp are measures of development. Higher the per capita income and gdpp better is the country's development. Income per capita and gdpp seems lowest for countries in clusters 0 and Child Mortality is highest for Cluster 0 Hence, these countries need some help.

```
In [40]: 1 country1[['child_mort', 'exports', 'health', 'income', 'gdpp', 'cluster_id']].groupby('cluster_id').mean().plot(kind='bar')
2         plt.show()
```



```
In [41]: 1 country1[['income', 'gdpp', 'child_mort', 'cluster_id']].groupby('cluster_id').mean().plot(kind='bar')
2         plt.show()
```



By using K= 3, Analysis of clusters formed from k mean on basis of 3 columns i.e income, gdpp and child_mort

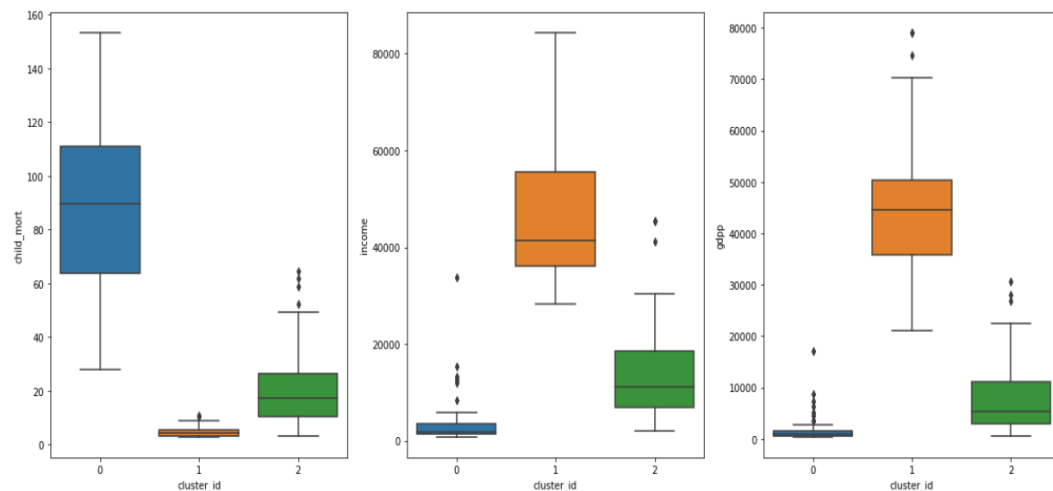
Inference:

Cluster 0 has lowest GDPP so we can say that the countries in cluster 0 must be in high Aid.

Cluster 0 has lowest income so we can say that the countries in cluster 0 must be in high Aid.

Cluster 0 has highest mortality rate so we can say that the countries in cluster 0 must be in high Aid.

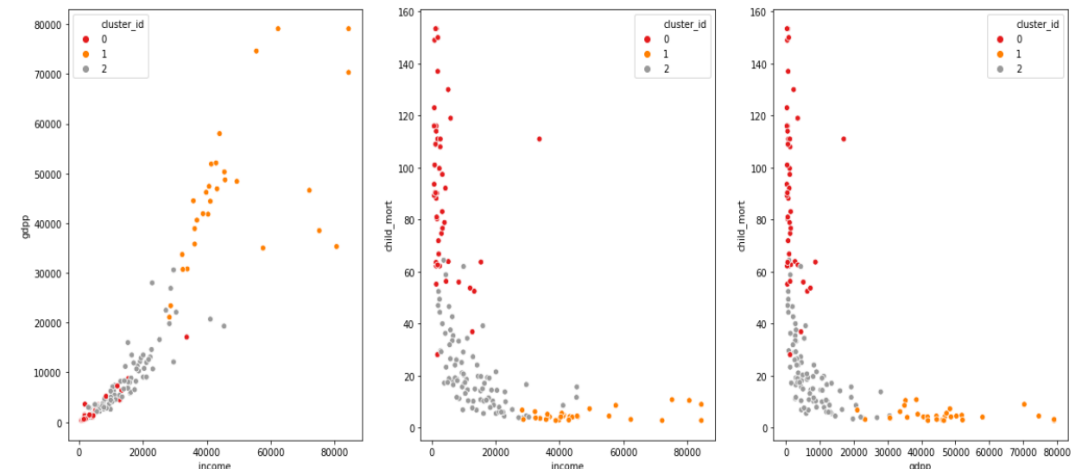
```
[ ]: 1 plt.figure(figsize=(20,8))
2     plt.subplot(1,3,1)
3     sns.boxplot(x='cluster_id', y='child_mort', data=country1)
4     plt.subplot(1,3,2)
5     sns.boxplot(x='cluster_id', y='income', data=country1)
6     plt.subplot(1,3,3)
7     sns.boxplot(x='cluster_id', y='gdpp', data=country1)
8     plt.show()
```



Final Inference:

Child Mortality is highest for Cluster 0, These clusters need some aid. Income and Gdpp are measures of development. Higher the per capita income and gdpp better is the country's development. Income per capita and gdpp seems lowest for countries in clusters 0. Hence, these countries need some help.

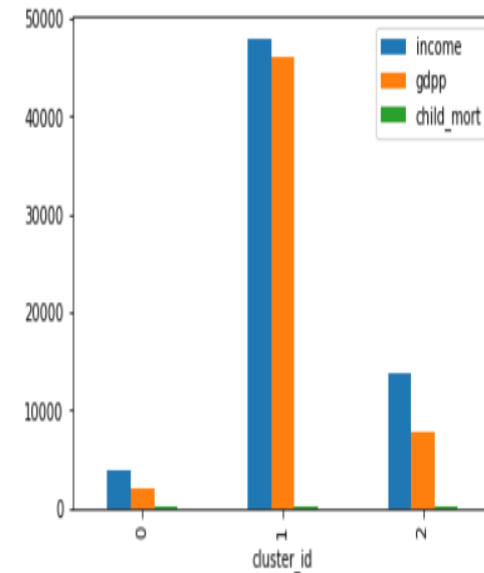
```
[35]: 1 # plotting the cluster ith respect to he clusters obtained
2     plt.figure(figsize=[20,8])
3     plt.subplot(1,3,1)
4     sns.scatterplot(x= 'income', y= 'gdpp', hue='cluster_id', legend='full', data=country1, palette= 'Set1')
5     plt.subplot(1,3,2)
6     sns.scatterplot(x= 'income', y= 'child_mort', hue='cluster_id', legend='full', data=country1, palette= 'Set1')
7     plt.subplot(1,3,3)
8     sns.scatterplot(x= 'gdpp', y= 'child_mort', hue='cluster_id', legend='full', data=country1, palette= 'Set1')
9     plt.show()
```



Cluster Profiling

	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
cluster_id									
0	90.335417	879.097657	115.348635	827.327888	3901.010000	10.608604	59.567083	4.972233	1911.400833
1	20.547778	3477.250726	528.925228	3589.291996	13804.333333	7.131624	73.393333	2.242591	7808.577778
2	4.989655	25405.359310	4253.879655	21316.695862	47784.413793	2.906731	80.453103	1.757352	46068.137931

```
1 country1[['income','gdpp','child_mort', 'cluster_id']].groupby('cluster_id').mean().plot(kind='bar')  
2 plt.show()
```



By Kmean clustering Top 10 countries

These countries are
sorted based on
ascending form of
income and gdpp and
descending form of
child_mortality

```
42]: 1 .groupby('cluster_id1').sort_values(by='child_mort', ascending= False)
      2 ster_country=country1[country1['cluster_id']==0].sort_values(by=['income','gdpp','child_mort'],ascending=[True, True,False])
      3 ster_country.head(10)
```

```
42]:
```

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	cluster_id
88	Liberia	89.3	62.457000	38.5860	302.80200	742.24	5.47	60.8	5.0200	331.62	0
37	Congo, Dem. Rep.	116.0	137.274000	26.4194	165.66400	742.24	20.80	57.5	6.5400	334.00	0
26	Burundi	93.6	22.243716	26.7960	104.90964	764.00	12.30	57.7	6.2600	331.62	0
112	Niger	123.0	77.256000	17.9568	170.86800	814.00	2.55	58.8	6.5636	348.00	0
31	Central African Republic	149.0	52.628000	17.7508	118.19000	888.00	2.01	47.5	5.2100	446.00	0
106	Mozambique	101.0	131.985000	21.8299	193.57800	918.00	7.64	54.5	5.5600	419.00	0
94	Malawi	90.5	104.652000	30.2481	160.19100	1030.00	12.10	53.1	5.3100	459.00	0
63	Guinea	109.0	196.344000	31.9464	279.93600	1190.00	16.10	58.0	5.3400	648.00	0
150	Togo	90.3	196.176000	37.3320	279.62400	1210.00	1.18	58.7	4.8700	488.00	0
132	Sierra Leone	153.4	67.032000	52.2690	137.65500	1220.00	17.20	55.0	5.2000	399.00	0

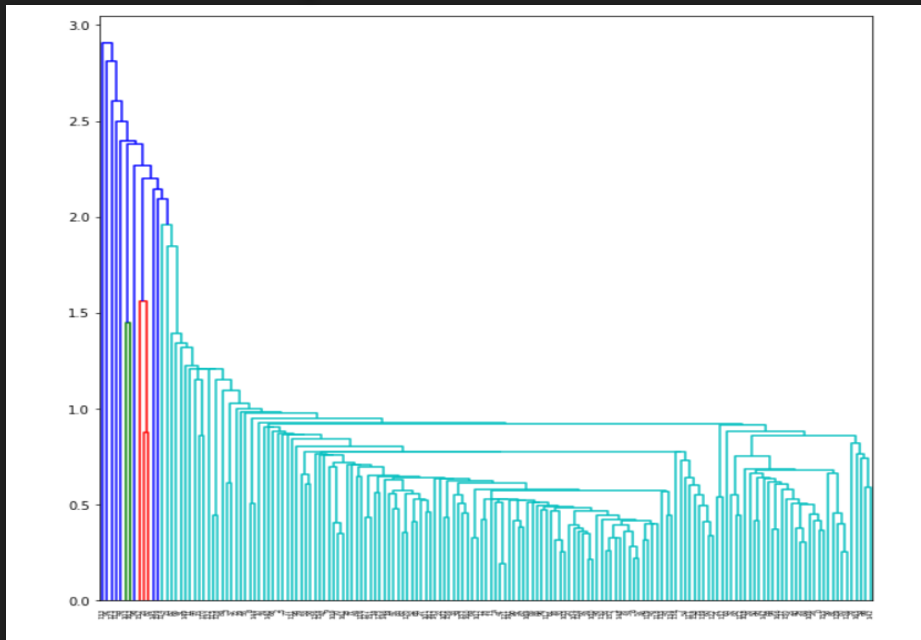
Hierarchal clustdring

Cluster made using cluster_labels variable

Types of hierarchal clustering

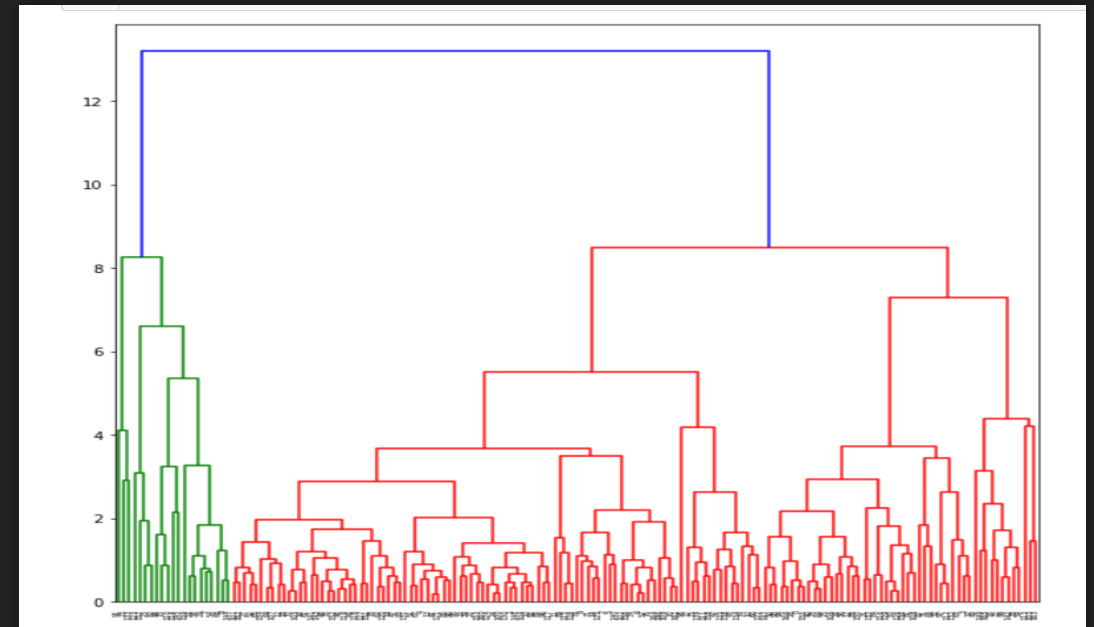
Single

○ Single Linkage do not give clear cluster formation so we have to try complete linkage in the next step.



Complete

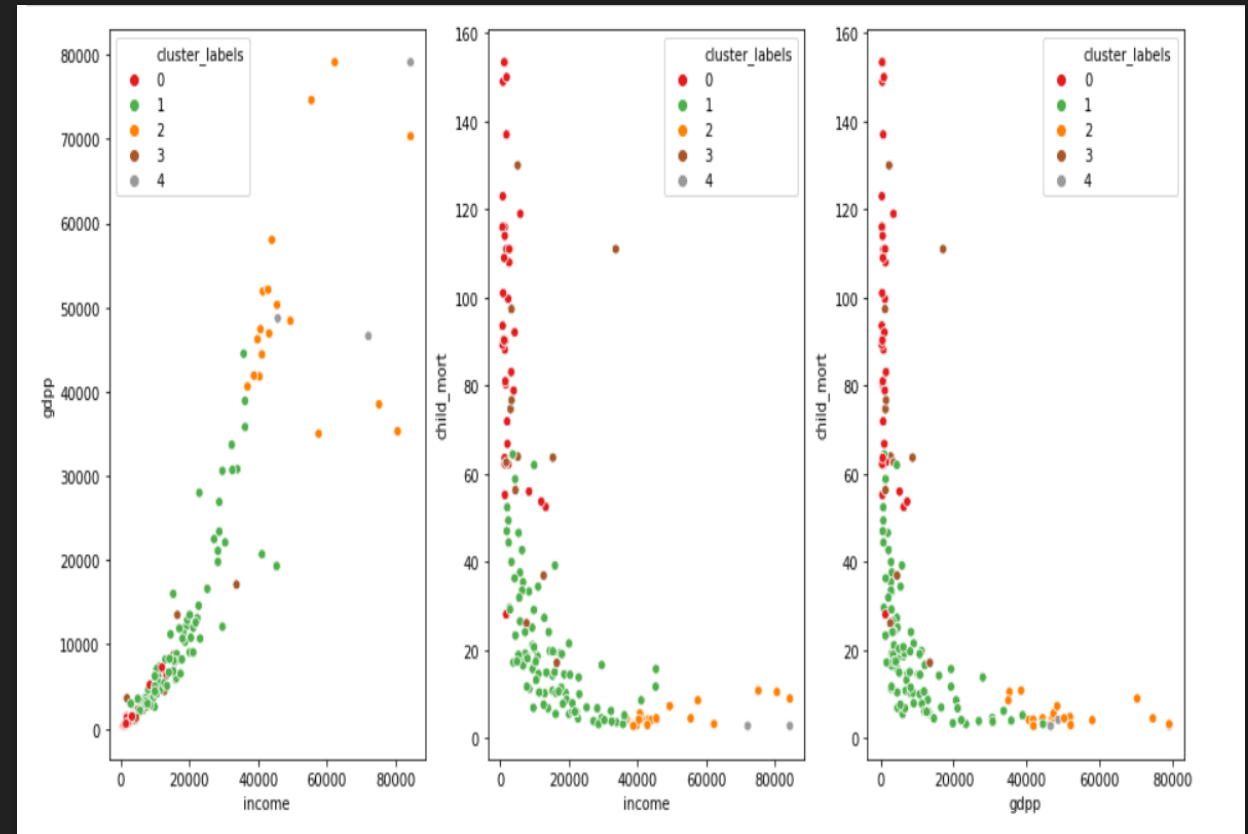
○ Now we see some good amount of clusters getting formed.



Choose cluster =4 from the dendrogram which is made by complete hirarichal clustering

Inference:

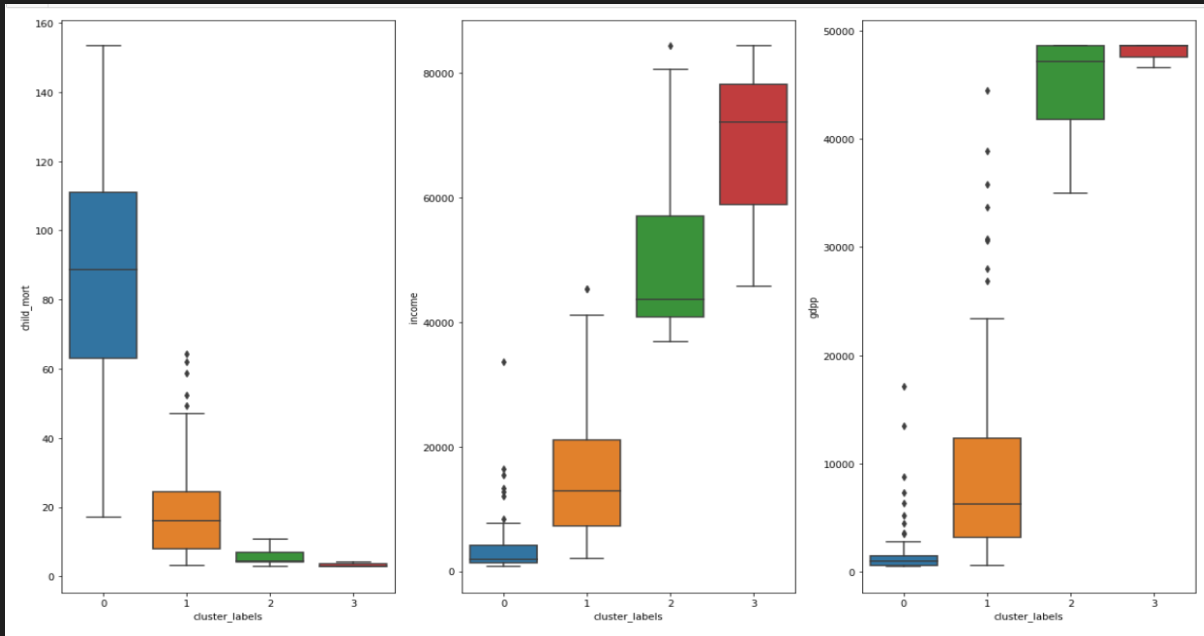
Child Mortality is highest for Cluster 0. These cluster need some aid. Income and Gdpp are measures of development. Higher the per capita income and gdpp better is the country's development. Income per capita and gdpp seems lowest for countries in clusters 0. Hence, these countries need some help.



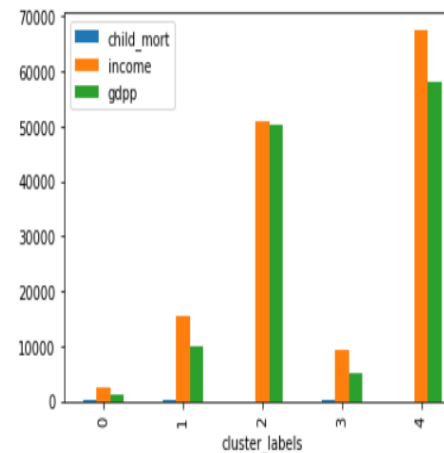
Cluster Profiling

Inference:

Child Mortality is highest for Cluster 0, These clusters need some aid. Income and Gdpp are measures of development. Higher the per capita income and gdpp better is the country's development. However Income per capita and gdpp seems lowest for countries in clusters 0. Hence, countries in cluster 0 need some help.



```
1 country1[['child_mort', 'income', 'gdpp', 'cluster_labels']].groupby('cluster_labels').mean().plot(kind='bar')  
2 plt.show()
```



By Hierarchical clustering method Top 10 countries

These countries are
sorted based on
ascending form of
income and gdpp and
descending form of
child_mortality

```
1 # let's filter the data with selected cluster
2 hirarichal_cluster_country = country1[country1['cluster_labels']==0].sort_values(by=['income', 'gdpp', 'child_mort'], ascending=[
3 hirarichal_cluster_country
```

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	cluster_labels
88	Liberia	89.3	62.457000	38.5860	302.80200	742.24	5.47	60.8	5.0200	331.62	0
37	Congo, Dem. Rep.	116.0	137.274000	26.4194	165.66400	742.24	20.80	57.5	6.5400	334.00	0
26	Burundi	93.6	22.243716	26.7960	104.90964	764.00	12.30	57.7	6.2600	331.62	0
112	Niger	123.0	77.256000	17.9568	170.86800	814.00	2.55	58.8	6.5636	348.00	0
31	Central African Republic	149.0	52.628000	17.7508	118.19000	888.00	2.01	47.5	5.2100	446.00	0
106	Mozambique	101.0	131.985000	21.8299	193.57800	918.00	7.64	54.5	5.5600	419.00	0
94	Malawi	90.5	104.652000	30.2481	160.19100	1030.00	12.10	53.1	5.3100	459.00	0
63	Guinea	109.0	196.344000	31.9464	279.93600	1190.00	16.10	58.0	5.3400	648.00	0
150	Togo	90.3	196.176000	37.3320	279.62400	1210.00	1.18	58.7	4.8700	488.00	0
132	Sierra Leone	153.4	67.032000	52.2690	137.65500	1220.00	17.20	55.0	5.2000	399.00	0

Top 10 countries

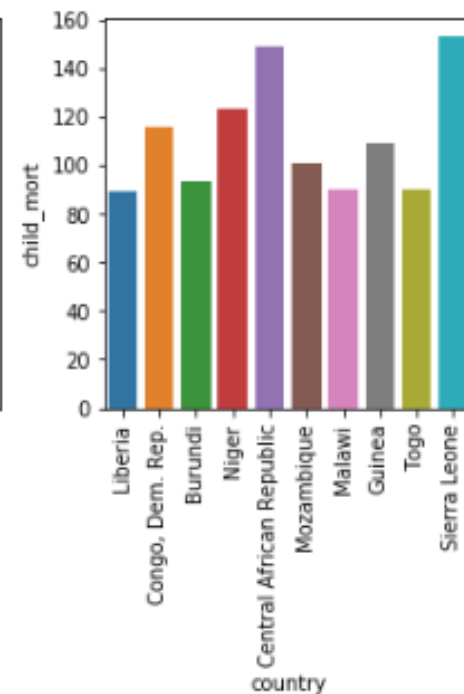
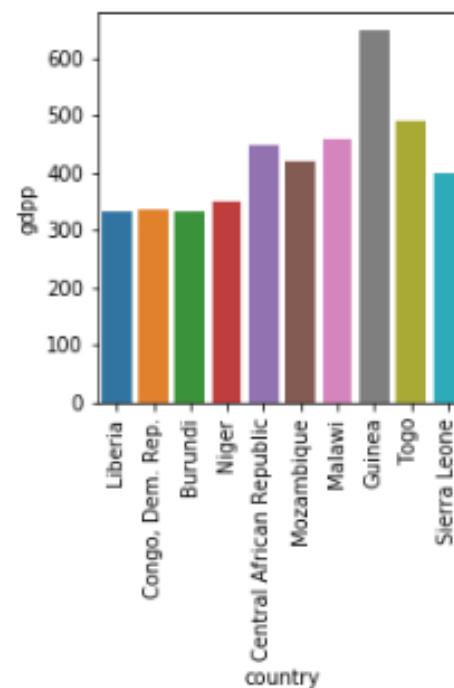
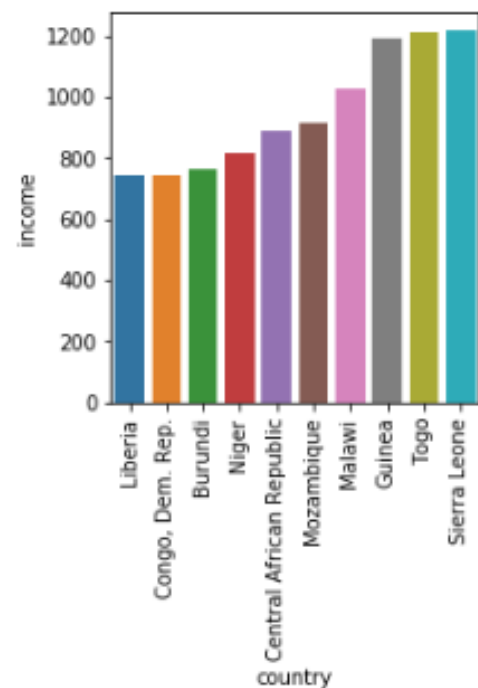
- 1) Liberia
- 2) Congo, Dem. Rep.
- 3) Burundi
- 4) Niger
- 5) Central African Republic
- 6) Mozambique
- 7) Malawi
- 8) Guinea
- 9) Togo
- 10) Sierra Leone

```
plt.subplot(1,3,1)
sns.barplot(x='country', y='income', data=select_count)
plt.xticks(rotation = 90)

plt.subplot(1,3,2)
sns.barplot(x='country', y='gdp', data=select_count)
plt.xticks(rotation = 90)

plt.subplot(1,3,3)
sns.barplot(x='country', y='child_mort', data=select_count)
plt.xticks(rotation = 90)

fig.tight_layout()
plt.show()
```



Final list of top 10 countries that needs Aid from CEO

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	cluster_labels1
88	Liberia	89.3	62.457000	38.586000	302.80200	742.24	5.47	60.8	5.0200	331.62	0
37	Congo, Dem. Rep.	116.0	137.274000	26.419400	165.66400	742.24	20.80	57.5	6.5400	334.00	0
26	Burundi	93.6	22.243716	26.796000	104.90964	764.00	12.30	57.7	6.2600	331.62	0
112	Niger	123.0	77.256000	22.243716	170.86800	814.00	2.55	58.8	6.5636	348.00	0
31	Central African Republic	149.0	52.628000	22.243716	118.19000	888.00	2.01	47.5	5.2100	446.00	0
106	Mozambique	101.0	131.985000	22.243716	193.57800	918.00	7.64	54.5	5.5600	419.00	0
94	Malawi	90.5	104.652000	30.248100	160.19100	1030.00	12.10	53.1	5.3100	459.00	0
63	Guinea	109.0	196.344000	31.946400	279.93600	1190.00	16.10	58.0	5.3400	648.00	0
150	Togo	90.3	196.176000	37.332000	279.62400	1210.00	1.18	58.7	4.8700	488.00	0
132	Sierra Leone	153.4	67.032000	52.269000	137.65500	1220.00	17.20	55.0	5.2000	399.00	0

Conclusion

- Based on my analysis I followed below observations to choose the countries that are in need of aid:
- Analysis based on : both K-means and Hierarchical clustering
- Cluster formation: identical
- I choose : k-mean clustering method and deduce the final list of countries which are in need of aid.

FINAL LIST OF TOP 10 COUNTRIES (USING K MEAN) TO FOCUS ON

```
42]: 1 .groupby('cluster_id1').sort_values(by='child_mort', ascending= False)
      2 cluster_country=country1[country1['cluster_id']==0].sort_values(by=['income','gdpp','child_mort'],ascending=[True, True,False])
      3 cluster_country.head(10)
```

```
42]:
```

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	cluster_id
88	Liberia	89.3	62.457000	38.5860	302.80200	742.24	5.47	60.8	5.0200	331.62	0
37	Congo, Dem. Rep.	116.0	137.274000	26.4194	165.66400	742.24	20.80	57.5	6.5400	334.00	0
26	Burundi	93.6	22.243716	26.7960	104.90964	764.00	12.30	57.7	6.2600	331.62	0
112	Niger	123.0	77.256000	17.9568	170.86800	814.00	2.55	58.8	6.5636	348.00	0
31	Central African Republic	149.0	52.628000	17.7508	118.19000	888.00	2.01	47.5	5.2100	446.00	0
106	Mozambique	101.0	131.985000	21.8299	193.57800	918.00	7.64	54.5	5.5600	419.00	0
94	Malawi	90.5	104.652000	30.2481	160.19100	1030.00	12.10	53.1	5.3100	459.00	0
63	Guinea	109.0	196.344000	31.9464	279.93600	1190.00	16.10	58.0	5.3400	648.00	0
150	Togo	90.3	196.176000	37.3320	279.62400	1210.00	1.18	58.7	4.8700	488.00	0
132	Sierra Leone	153.4	67.032000	52.2690	137.65500	1220.00	17.20	55.0	5.2000	399.00	0



THANK YOU