CLUSTERING ASSIGNMENT

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.

The objective of this assignment is to categorize the countries using some socio-economic and health factors that determine the overall development of the country and to suggest the countries which needs more focus.

SUBMITTED BY :

Aruna D

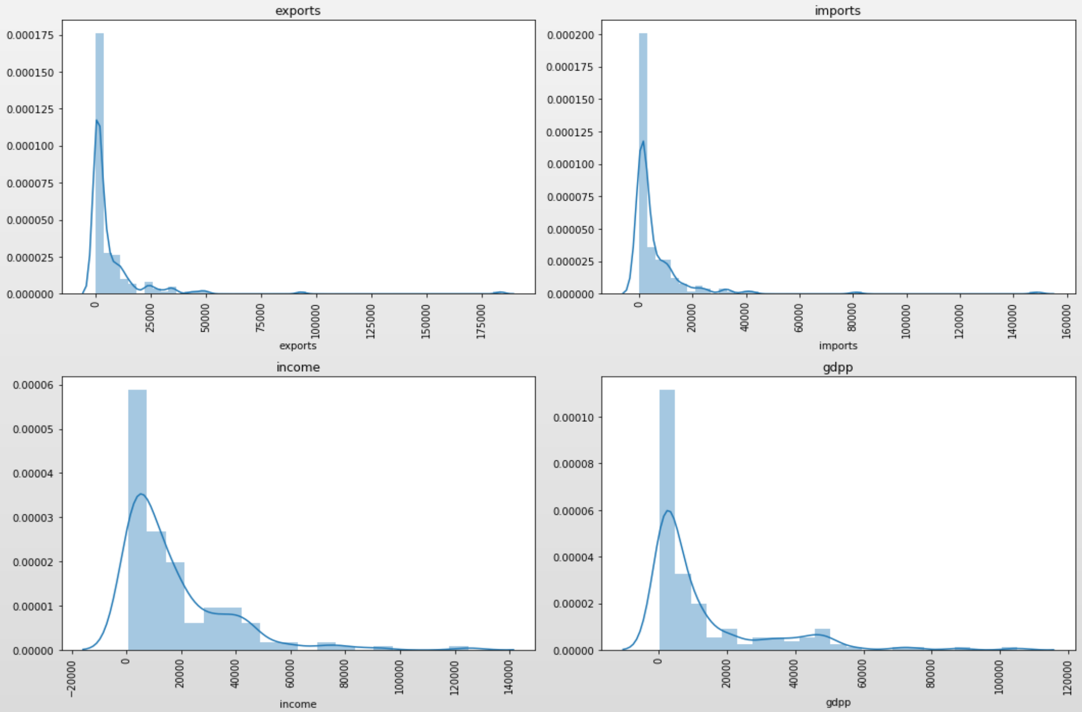
STEPS FOLLOWED :

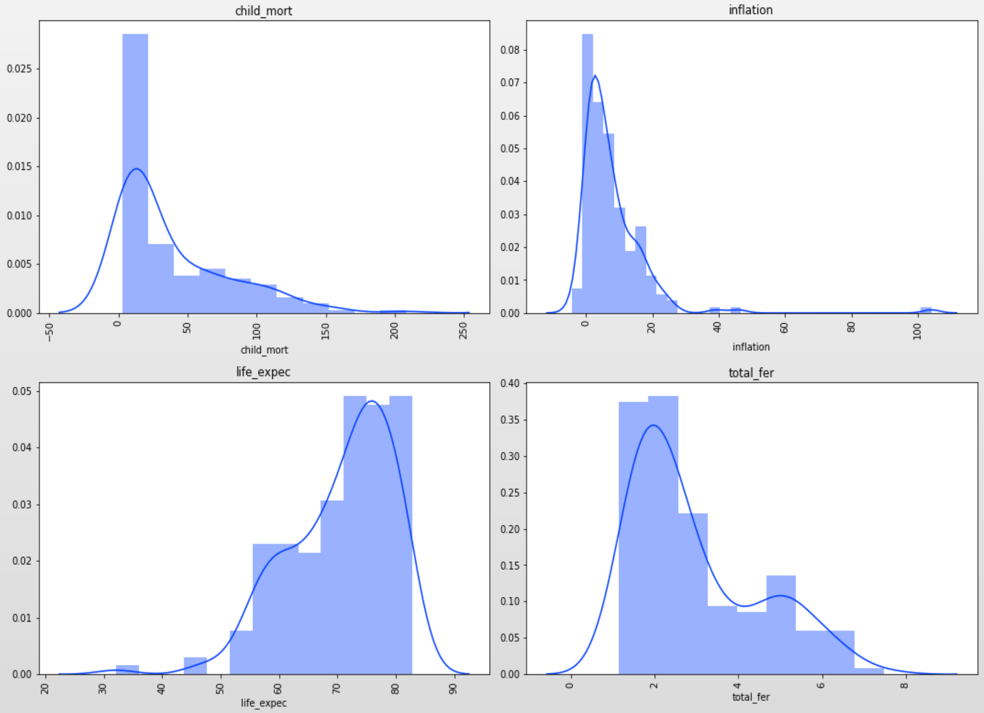
- Data Reading
- Data preparation and Data understanding :
  - Routine data check
  - Data quality check, missing value analysis and duplicate rows check
  - Converting export, health, import which is given as percentage of gdp to their respective actual values.

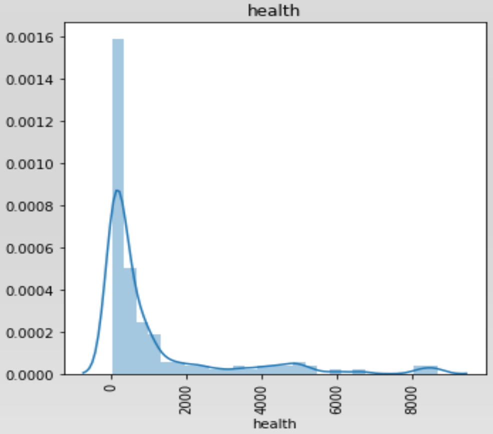
There was no missing values or duplicated row in data set

- Data Visualisation and EDA:
  - In EDA, Univariate analysis, Bivariate analysis, Multi variate analysis was done.
- Outlier treatment and Hopkin's check
  - In outlier treatment, for child mortality the higher range outliers are not capped. For remaining columns lower range outliers are not capped.
- Scaling .
- K-Means Clustering (k=3)
- Hierarchical Clustering (number of clusters chosen is 3)
- Country identification.

EDA-UNIVARIATE ANALYSIS







Inferences :

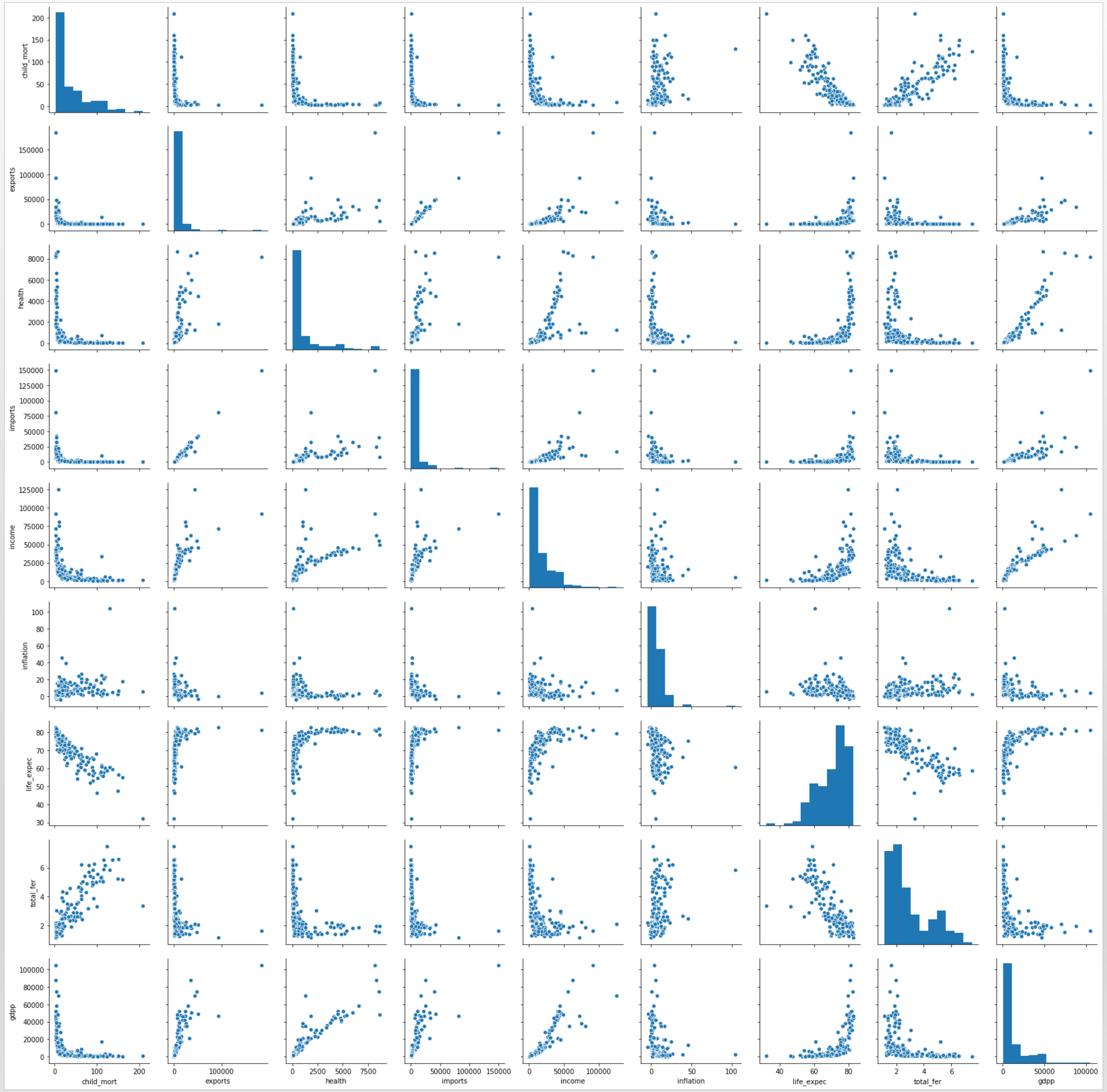
1. The average life expentancy is around 70 years.
2. The mean total_fer is around 3.

BIVARIATE ANALYIS:

Inferences from pair plot :

Following variables have linear relationship:

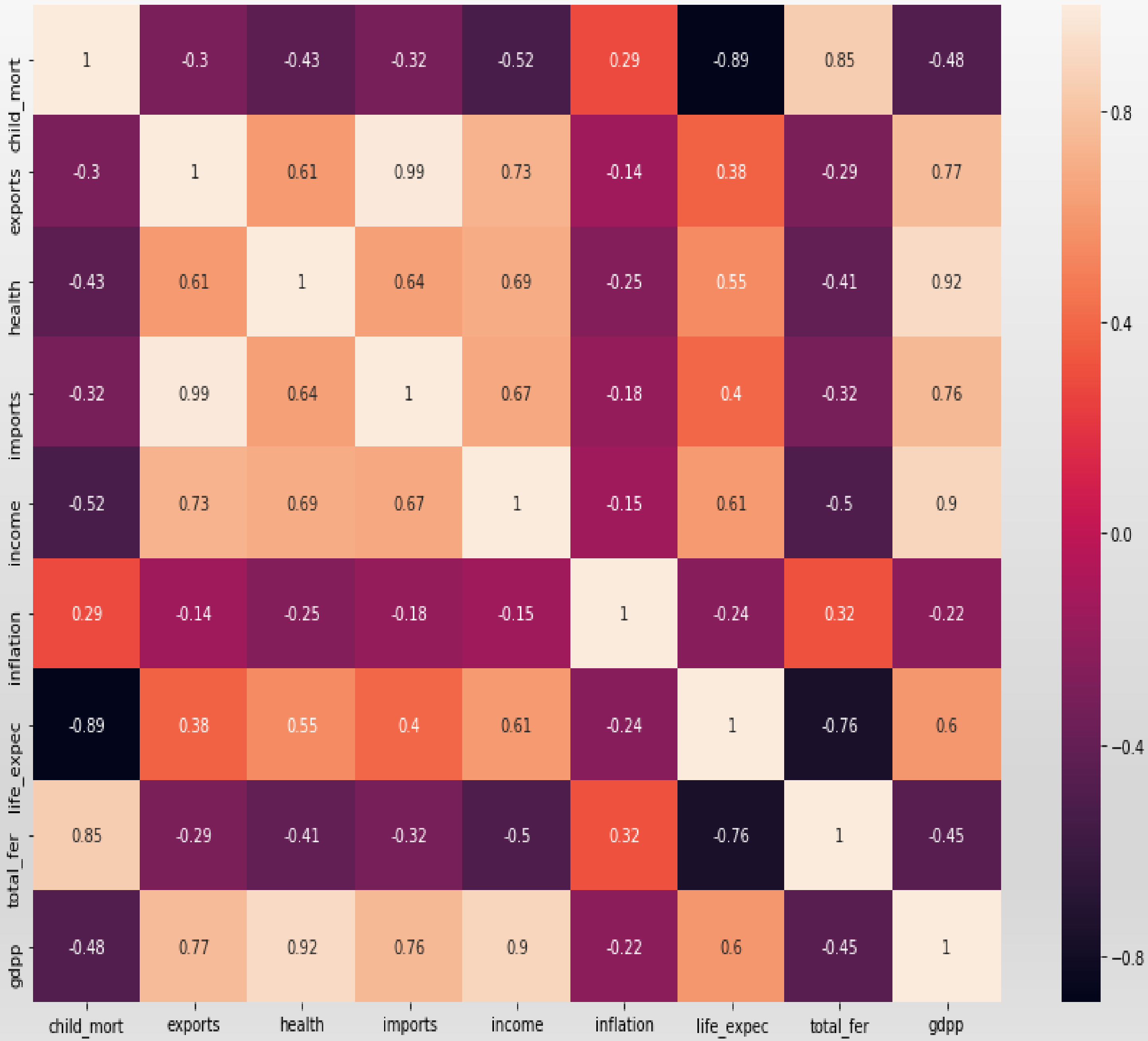
1. Child mortality and life expentency,
2. Child mortality and total_fert
3. Exports and imports
4. Health and GDP



MULTIVARIATE ANALYIS:

Inferences:

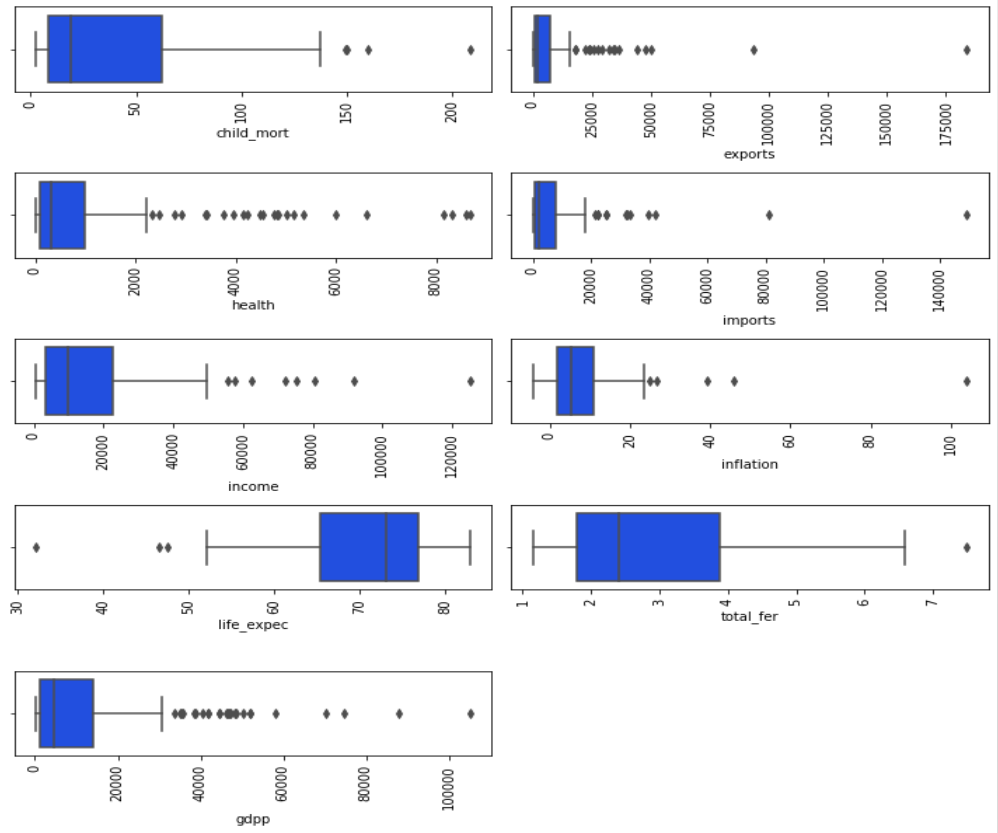
- Child mortality is strongly negatively correlated with life expectancy
- Child mortality is negatively correlated with GDP,income also
- Health is strongly positively correlated with GDP.
- We can see that GDP and health,income , exports,imports,life expentency are positively correlated and child mortality,inflation,total fertility are negatively correlated.
- Child mortality is highly positively correlated with total fertility.
- Exports and imports have highest positive correlation
- **So we can conclude that, Underdeveloped countries will have low GDP.As GDP and health are highly correlated, health will be bad and child mortality will be high.Income will also be low for underdeveloped countries**



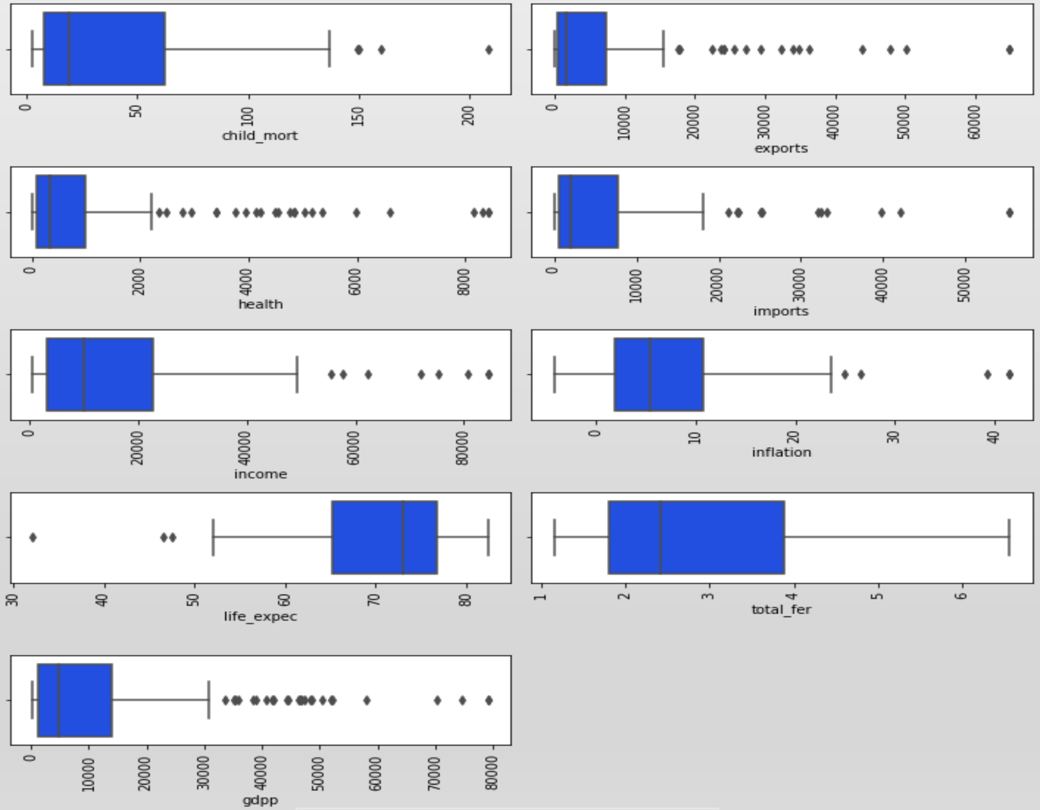
	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
child_mort	1	-0.3	-0.43	-0.32	-0.52	0.29	-0.89	0.85	-0.48
exports	-0.3	1	0.61	0.99	0.73	-0.14	0.38	-0.29	0.77
health	-0.43	0.61	1	0.64	0.69	-0.25	0.55	-0.41	0.92
imports	-0.32	0.99	0.64	1	0.67	-0.18	0.4	-0.32	0.76
income	-0.52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

OUTLIER TREATMENT

In outlier treatment, for child mortality the higher range outliers are not capped. For remaining columns lower range outliers are not capped. This is because if we cap higher outlier range for child mortality, we will lose the information about country which are in real need. Similarly for Income, GDP and all low values if capped, will lead to lose of data about under developed countries.



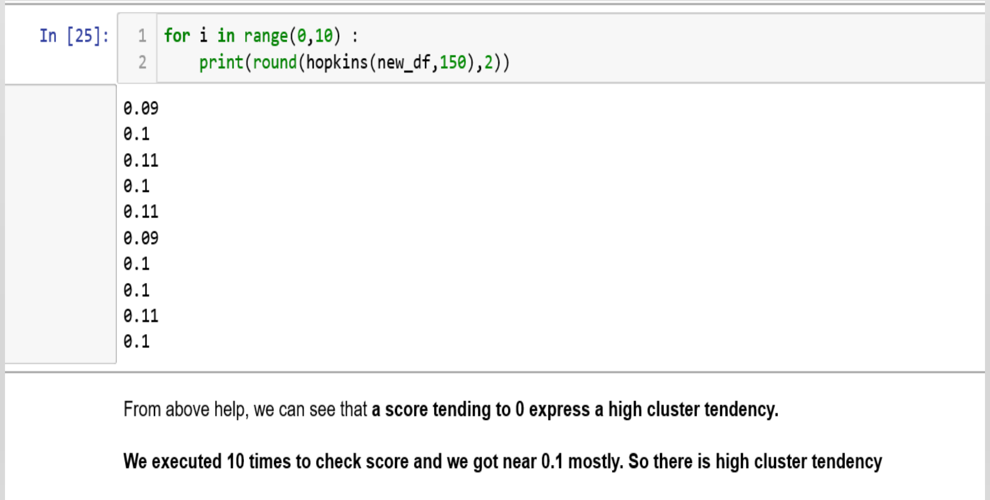
Before outlier treatment



After outlier treatment

HOPKINS TEST

- METHOD-1
- Used “Hopkins”function from external library pyclustertend
- hopkins(data_frame, sampling_size) Assess the clusterability of a dataset. A score between 0 and 1, a score around 0.5 express no clusterability and a score tending to 0 express a high cluster tendency.



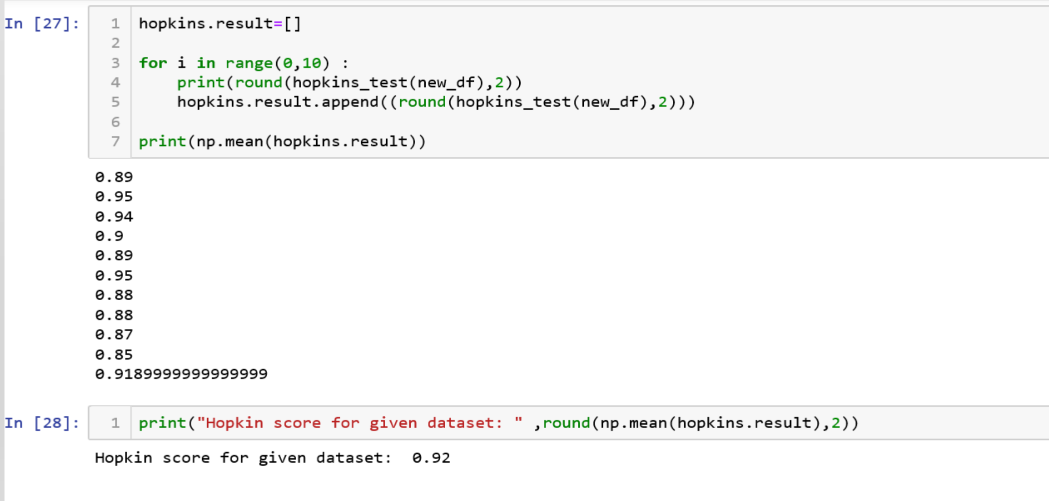
```
In [25]:
for i in range(0,10) :
    print(round(hopkins(new_df,150),2))

0.09
0.1
0.11
0.1
0.11
0.09
0.1
0.1
0.11
0.1
```

From above help, we can see that **a score tending to 0 express a high cluster tendency.**

We executed 10 times to check score and we got near 0.1 mostly. So there is high cluster tendency

USING FUNCTION: METHOD-2

The hopkins statistic (introduced by Brian Hopkins and John Gordon Skellam) is a way of measuring the cluster tendency of a dataset. It acts as a statistical hypothesis test where the null hypothesis is that the data is generated by a Poisson point process and are thus uniformly randomly distributed. A value close to 1 tends to indicate the data is highly clustered, random data will tend to result in values around 0.5, and uniformly distributed data will tend to result in values close to 0.



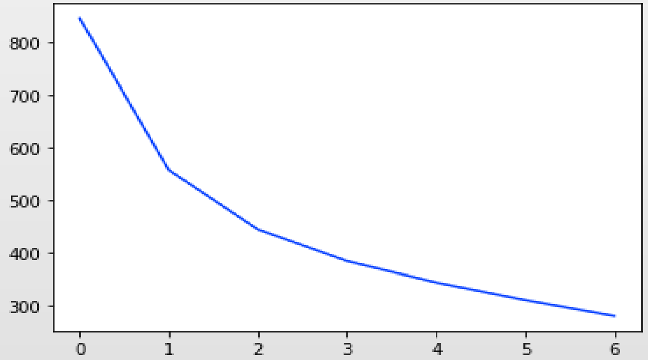
```
In [27]:
hopkins.result=[]

for i in range(0,10) :
    print(round(hopkins_test(new_df),2))
    hopkins.result.append((round(hopkins_test(new_df),2)))

print(np.mean(hopkins.result))

0.89
0.95
0.94
0.9
0.89
0.95
0.88
0.88
0.87
0.85
0.9189999999999999

In [28]:
print("Hopkin score for given dataset: " ,round(np.mean(hopkins.result),2))

Hopkin score for given dataset:  0.92
```

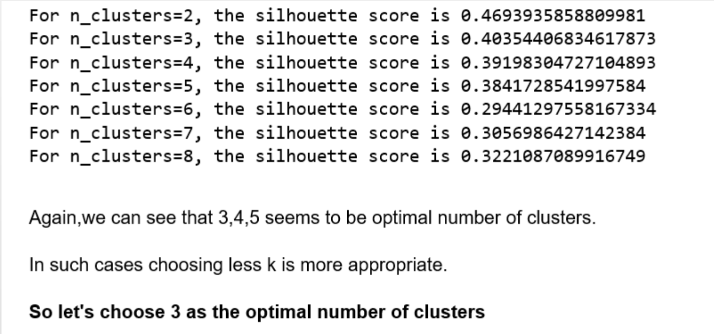
From both the above method, we got good hopkins score. So there is high cluster tendency in the given dataset

K-MEANS CLUSTERING:

We would perform following steps :

- ❑ Finding the Optimal Number of Clusters using both Elbow and Silhouette score





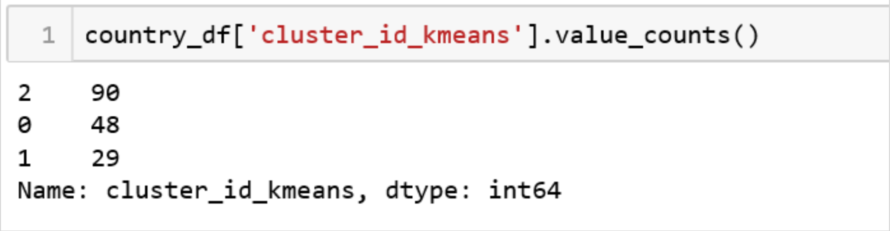
```
For n_clusters=2, the silhouette score is 0.4693935858809981
For n_clusters=3, the silhouette score is 0.40354406834617873
For n_clusters=4, the silhouette score is 0.39198304727104893
For n_clusters=5, the silhouette score is 0.3841728541997584
For n_clusters=6, the silhouette score is 0.29441297558167334
For n_clusters=7, the silhouette score is 0.3056986427142384
For n_clusters=8, the silhouette score is 0.3221087089916749
```

Again,we can see that 3,4,5 seems to be optimal number of clusters.

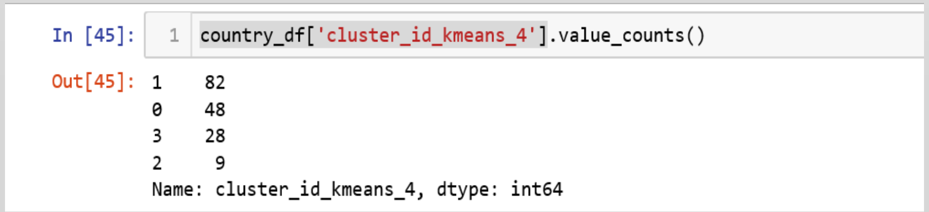
In such cases choosing less k is more appropriate.

So let's choose 3 as the optimal number of clusters

- ❖ 1st image represents Elbow curve/SSD curve. 2nd Image represents the Silhouette score. We can see that 3,4,5 are optimal number of clusters. It always good to choose smaller K in that case. So K can be 3 or 4.
- ❖ But for both K=3,K=4 the number of underdeveloped countries was 48. So let's choose K=3 itself.



```
country_df['cluster_id_kmeans'].value_counts()

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



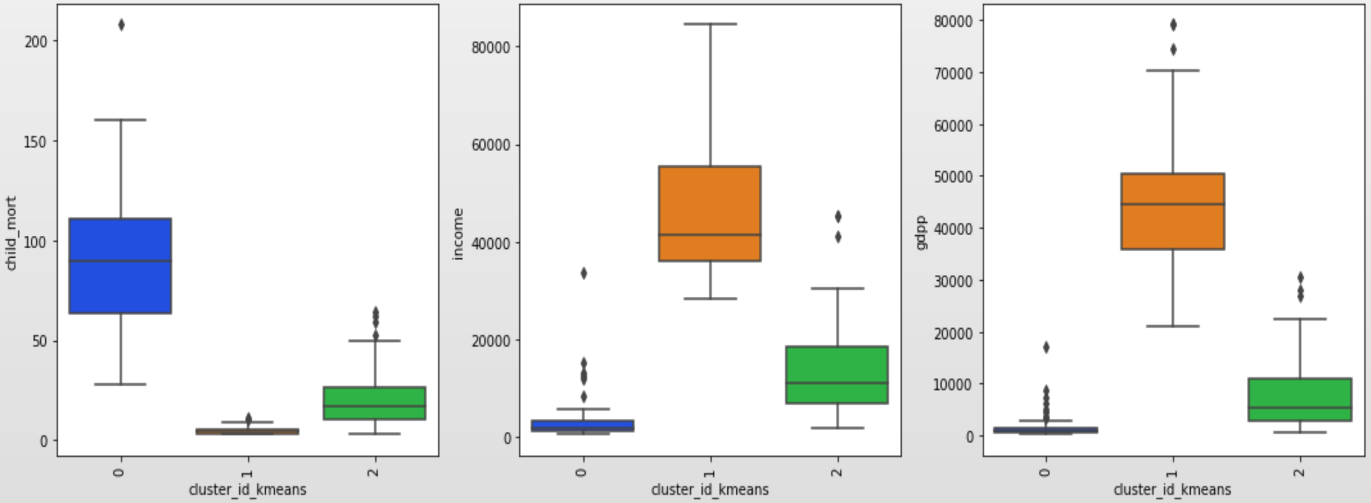
```
In [45]:  country_df['cluster_id_kmeans_4'].value_counts()

Out[45]: 1    82
         0    48
         3    28
         2     9
         Name: cluster_id_kmeans_4, dtype: int64
```

- ❑ Run K-Means with the chosen K

K-MEANS CLUSTERING (CONT.):

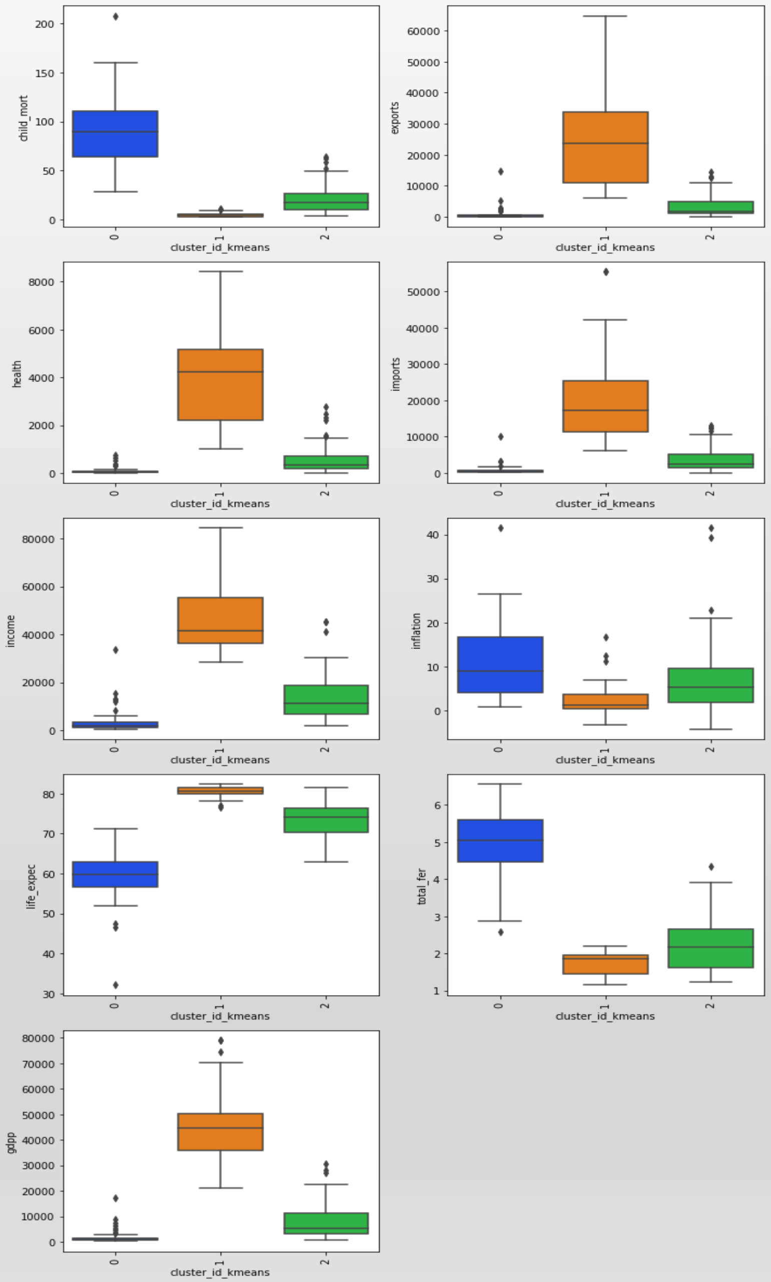
- Visualize the cluster.(right side graph)

- Cluster Profiling using "gdpp,child_mort and income“



- We can clearly see that:
- cluster 0 has highest child mortality, lowest income and lowest GDP
- cluster 1 has lowest child mortality, highest income and highest GDP
- cluster 2 has average child mortality, income and GDP
- **So we must focus on CLUSTER-0. Cluster-0 are under-developed countries that are in the direst need of aid.**

- Listing underdeveloped countries that are in the direst need of aid.(Around 48 countries)

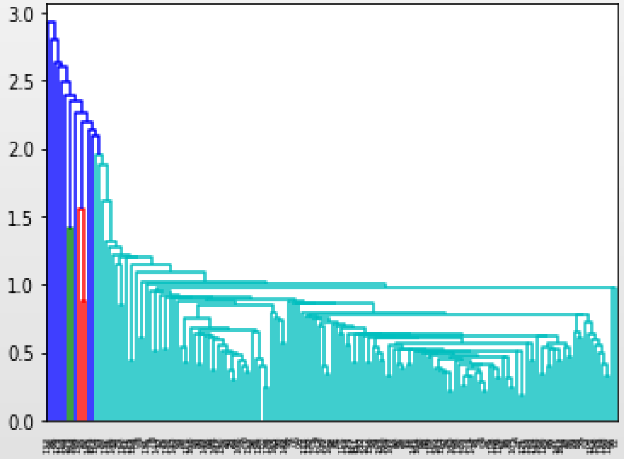


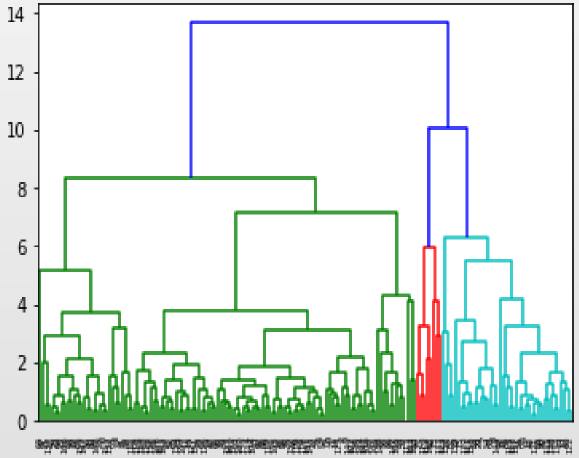
HIERARCHICAL CLUSTERING

We will perform below mentioned steps:

- ❑ We will use both single and complete linkage.
- ❑ Choose one method based on the results :

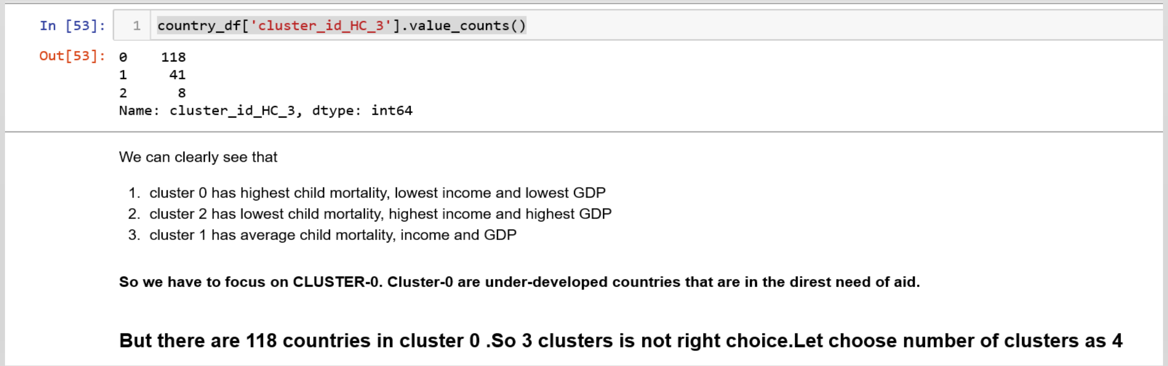
  1st image represents single linkage and 2nd image represents complete linkage. It's very clear that complete linkage does better clustering





- ❑ Number of clusters is chosen as 4. Because for 3 clusters, around 118 countries got clustered in underdeveloped countries cluster

.



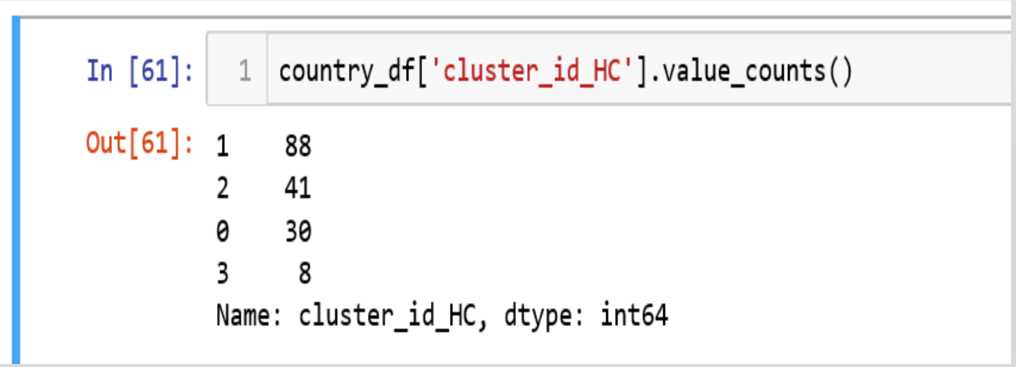
```
In [53]:   country_df['cluster_id_HC_3'].value_counts()
Out[53]: 0    118
         1     41
         2      8
         Name: cluster_id_HC_3, dtype: int64
```

We can clearly see that

1. cluster 0 has highest child mortality, lowest income and lowest GDP
2. cluster 2 has lowest child mortality, highest income and highest GDP
3. cluster 1 has average child mortality, income and GDP

So we have to focus on CLUSTER-0. Cluster-0 are under-developed countries that are in the direst need of aid.

But there are 118 countries in cluster 0 .So 3 clusters is not right choice.Let choose number of clusters as 4



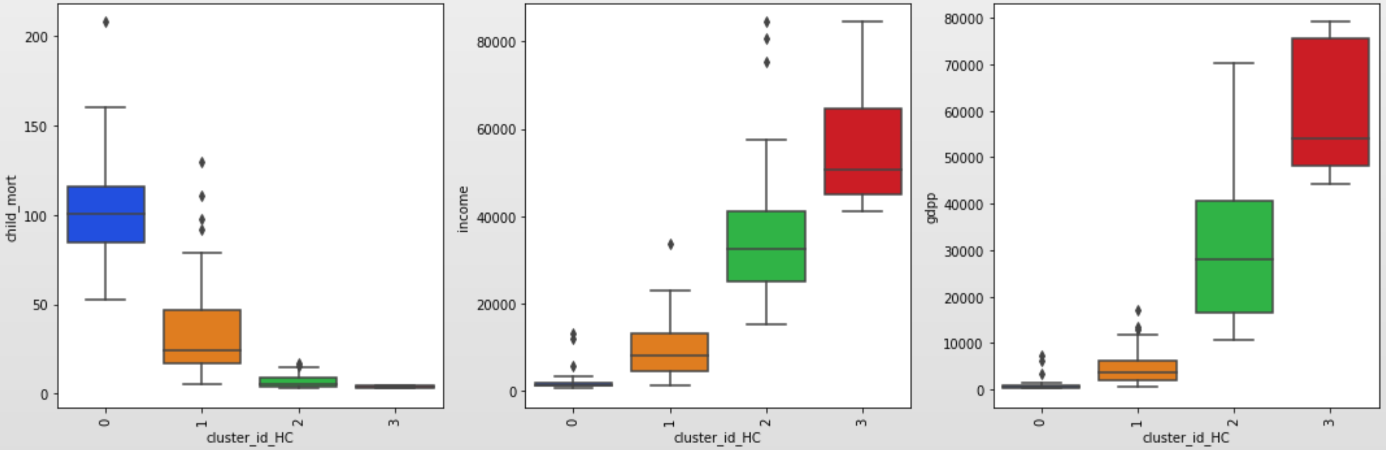
```
In [61]:   country_df['cluster_id_HC'].value_counts()
Out[61]: 1    88
         2    41
         0    30
         3     8
         Name: cluster_id_HC, dtype: int64
```

1st image represents the count of number of countries in each cluster when no of clusters=3 and 2nd image when no of clusters=4

HIERARCHICAL CLUSTERING(CONT.)

- ❑ Visualize the cluster.(right side graph)

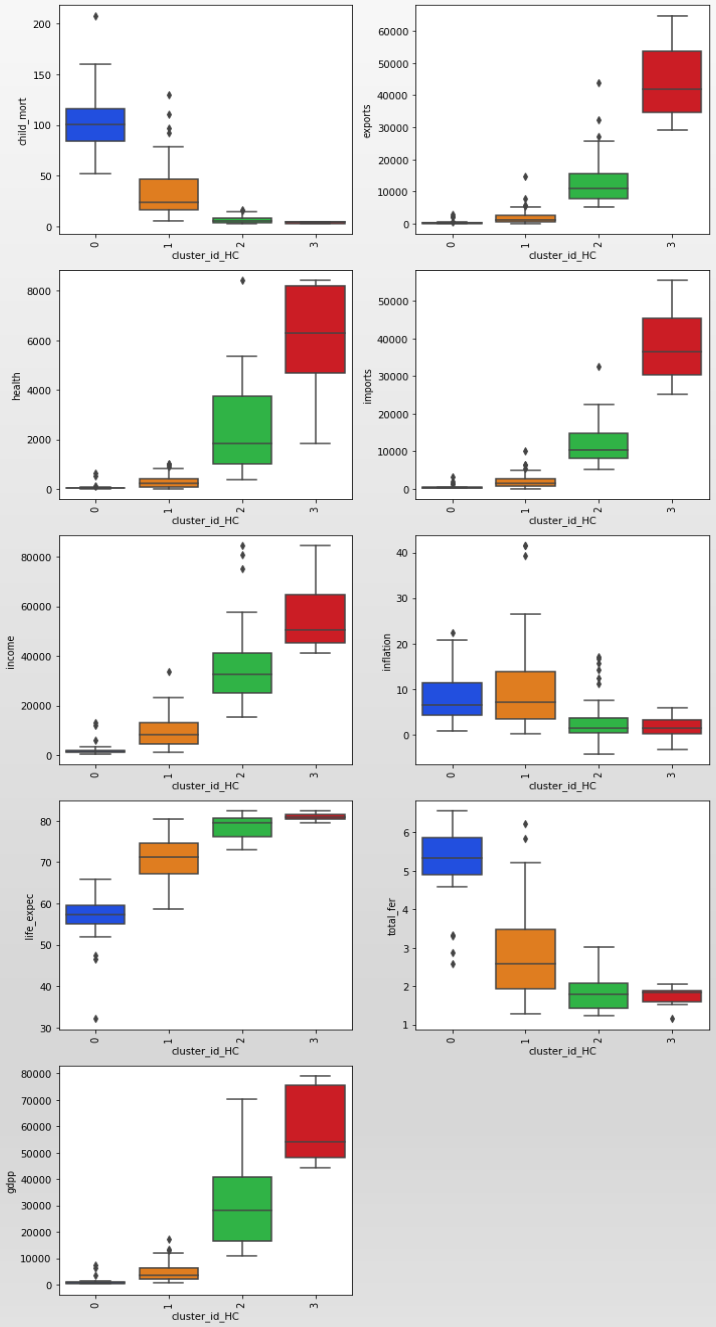
- ❑ Cluster Profiling using "gdpp,child_mort and income“



We can clearly see that:

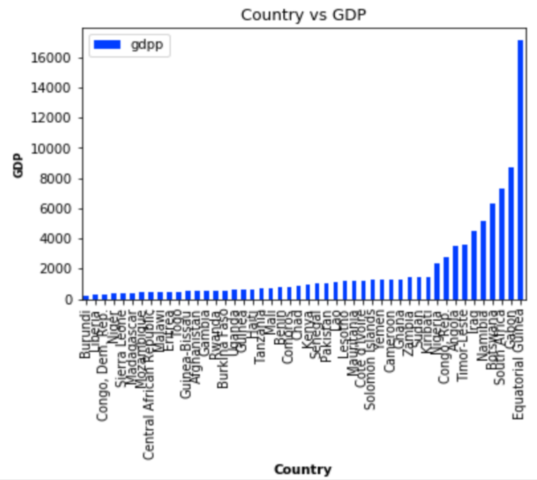
- ❖ cluster 0 has highest child mortality, lowest income and lowest GDP
- ❖ cluster 1 has 2nd high child mortality, low income and low GDP
- ❖ cluster 2 has low child mortality, high income and high GDP
- ❖ cluster 3 has lowest child mortality, highest income and highest GDP
- ❖ **So we must focus on CLUSTER-0. Cluster-0 are under-developed countries that are in the direst need of aid.**

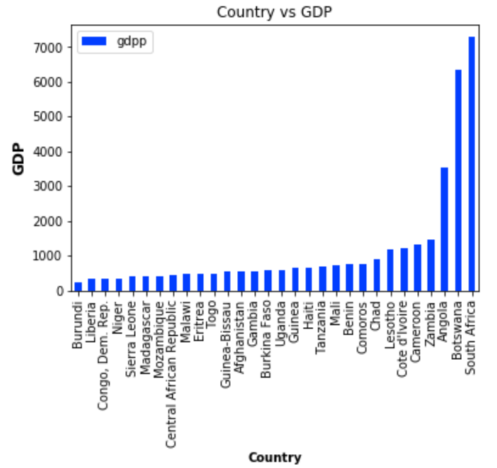
- ❑ Listing underdeveloped countries that are in the direst need of aid.(Around 30 countries)

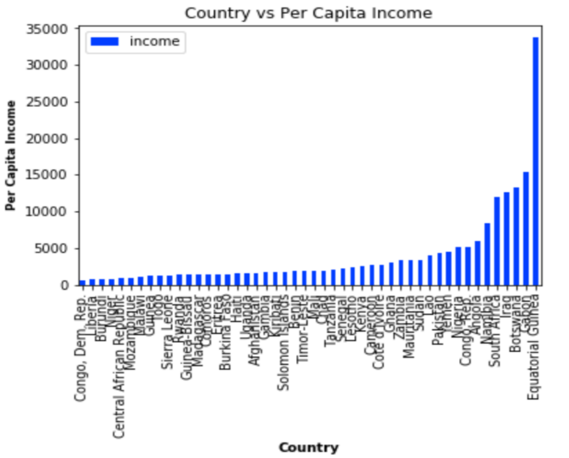


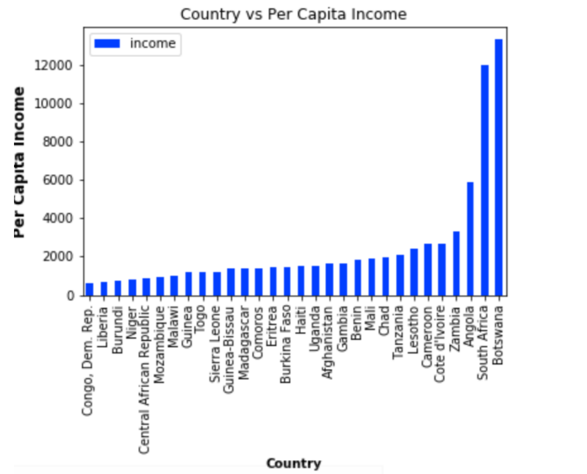
K-MEANS COUNTRY LIST VS HIERARCHICAL CLUSTERING(HC) COUNTRY LIST

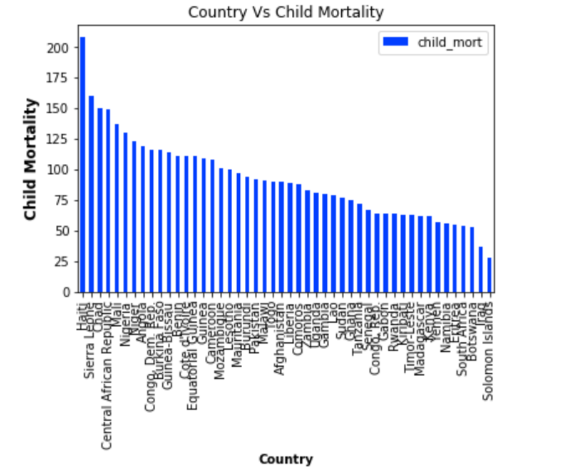
1^{st} , 2^{nd} ,3^{rd} graphs represents the GDP, child mortality, income of countries formed by k-means and hc respectively.

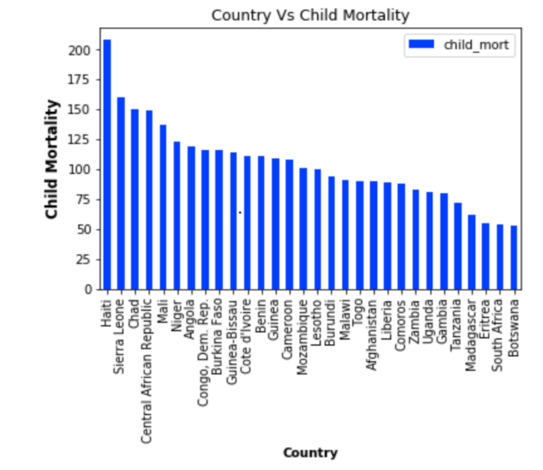










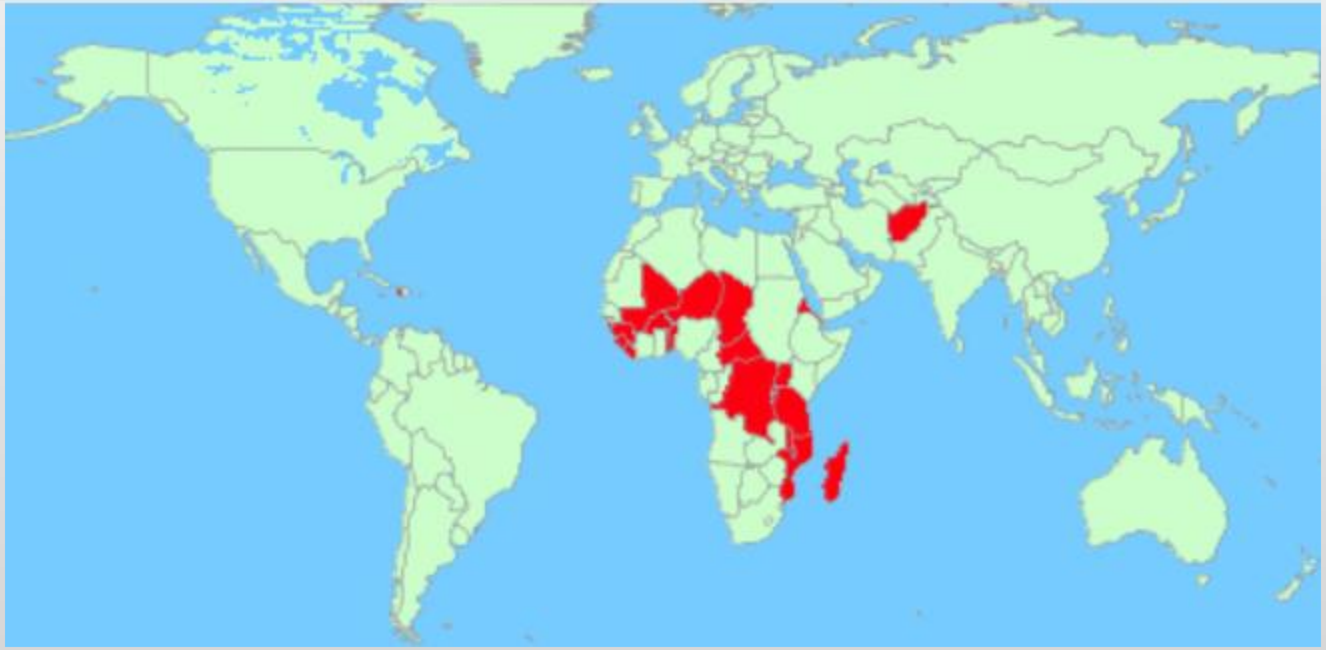


FINAL COUNTRY LIST TO FOCUS

From all the graphs , the final countries are as follows:

The countries are sorted from lowest to highest GDP.(These countries are marked in red color in world map.)

1.Burundi
2.Liberia
3.Congo, Dem. Rep.
4.Niger
5.Sierra Leone
6.Madagascar
7.Mozambique
8.Central African Republic
9.Malawi
10.Eritrea
11.Togo
12.Guinea-Bissau
13.Afghanistan
14.Gambia
15.Rwanda
16.Burkina Faso
17.Uganda
18.Guinea
19.Haiti
20.Tanzania
21.Mali
22.Benin
23.Comoros
24.Chad



THANK YOU