CLUSTERING OF COUNTRIES ASSIGNMENT

By:

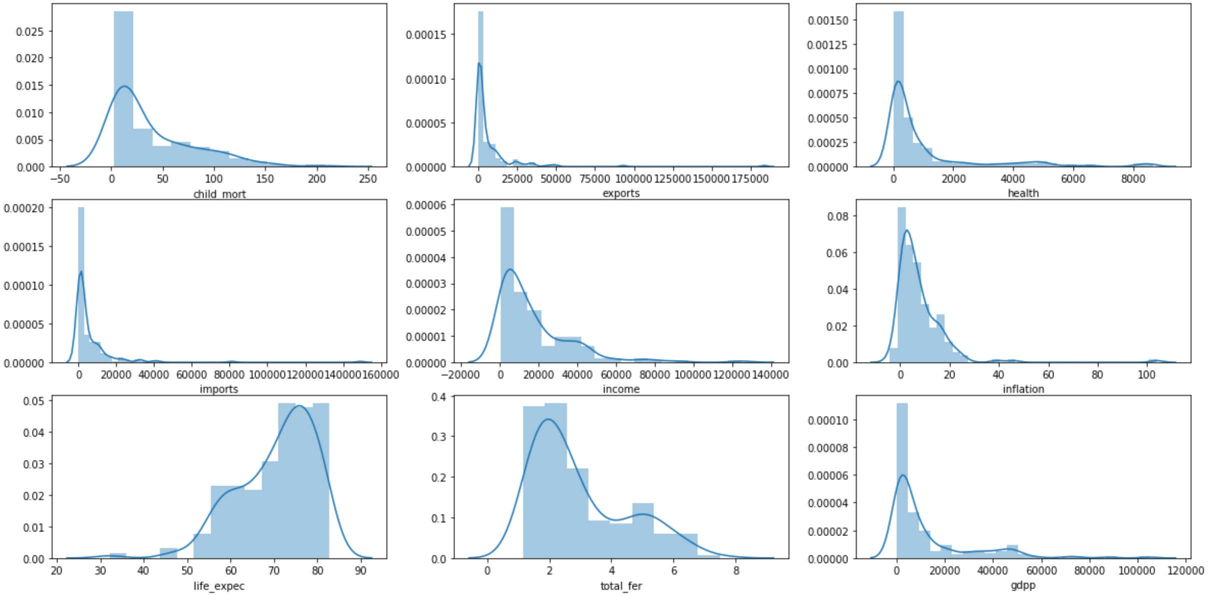
Mamatha E

Problem Statement

Categorizing the countries using socio-economic and health factors that determine the overall development of the country and identifying the top5 countries that are of immediate aid.

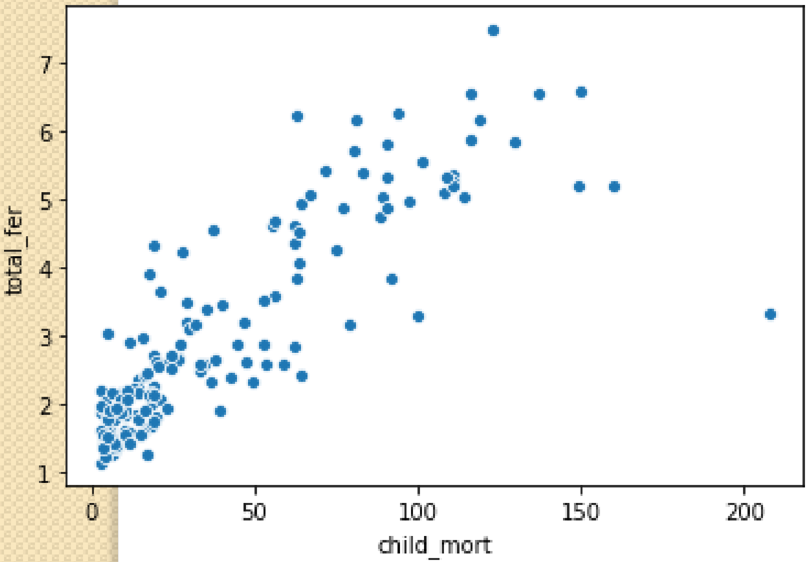
Methodology & Analysis

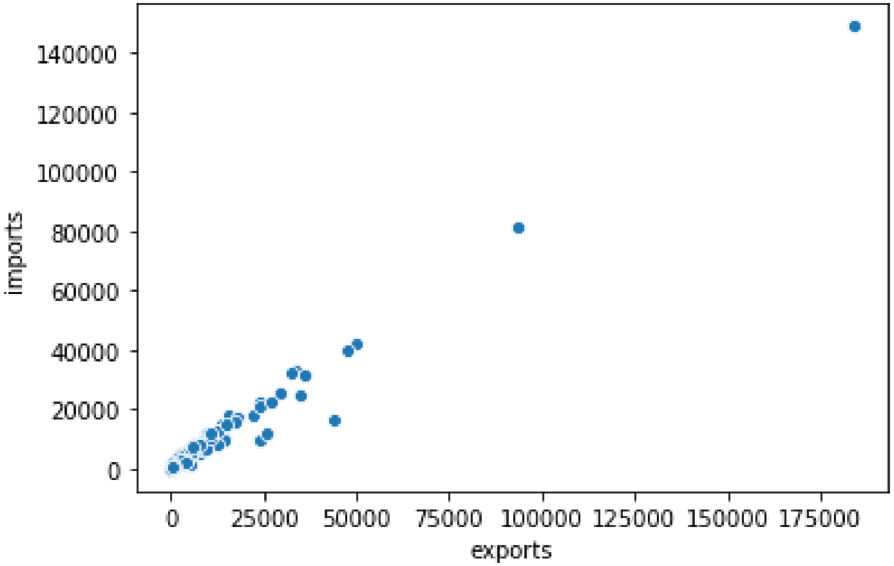
- Initially loading the data set and understanding it.
- Data cleaning was carried out, certain attribute values need to be updated.
- **EDA**:
  - ***UNIVARIATE ANALYSIS (CONTINUOUS):***
    - From the plots shown below, all the columns are skewed towards left except for the column ‘life_expec’ which is right skewed.



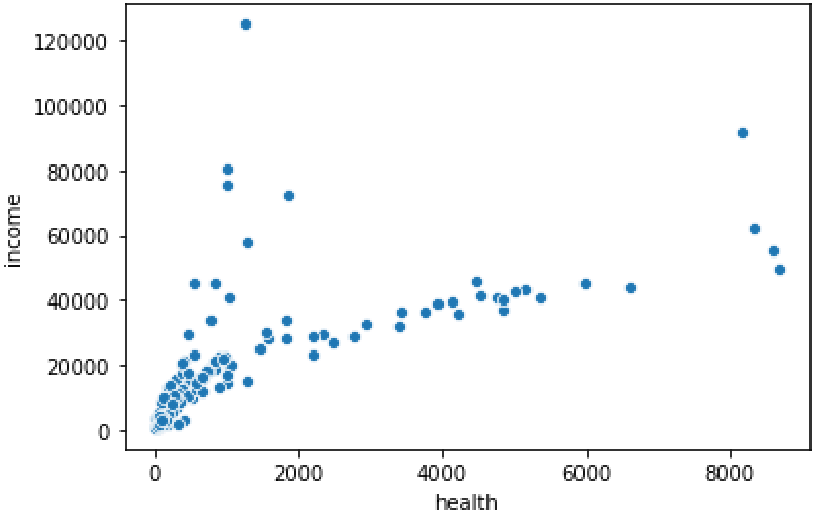
BIVARIATE ANALYSIS (CONTINUOUS-CONTINUOUS):

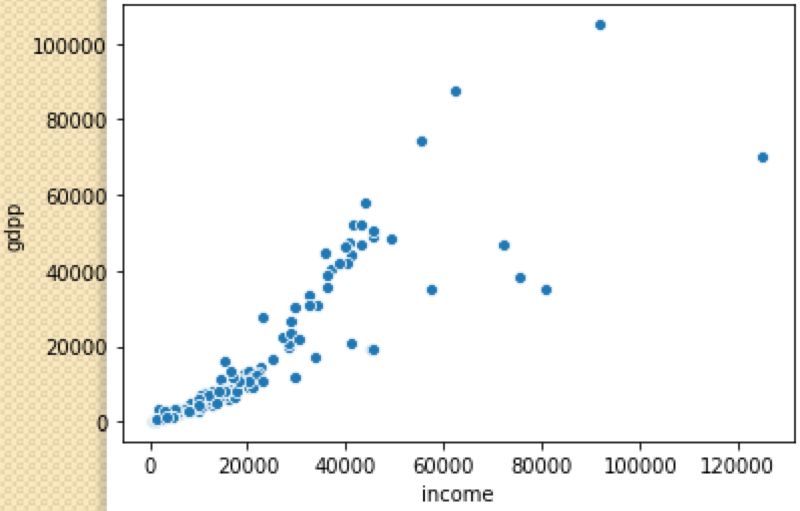
- Some of the variables are having correlation with other variables as follows:
- **Positive Correlation:**
  - *child_mort & total_fer*: As the death of infants are increasing, the infants are born at the same rate.
  - *exports & imports*: Exports and imports are also increasing at the same rate.

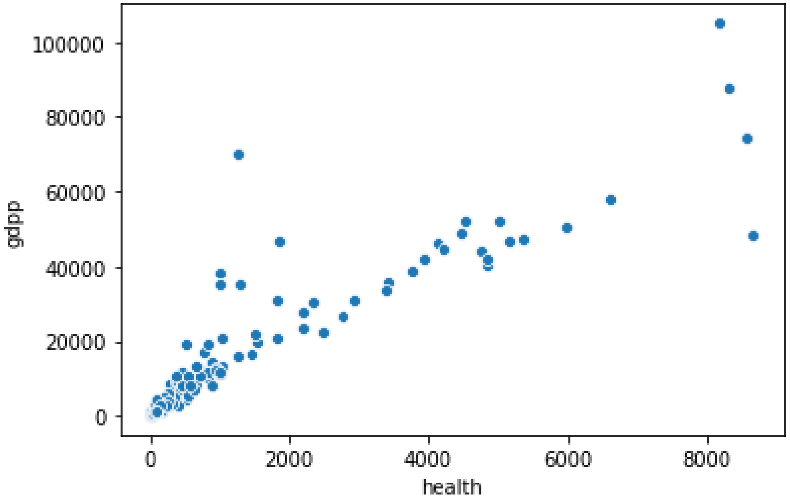




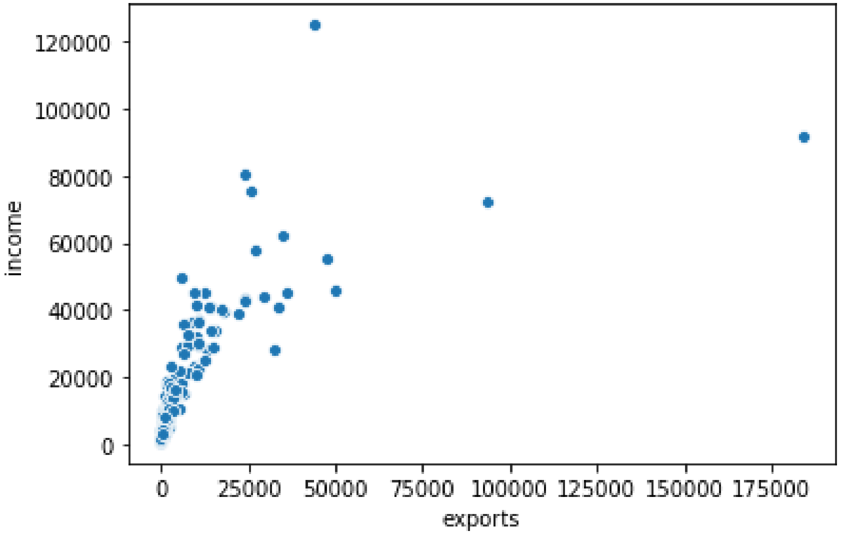
- *health & income, health & gdpp*: As the net income or gdpp increases , it implies that there is increase in health expense.
- *income & gdpp*: As the net income of a person are increasing so does the gdpp.

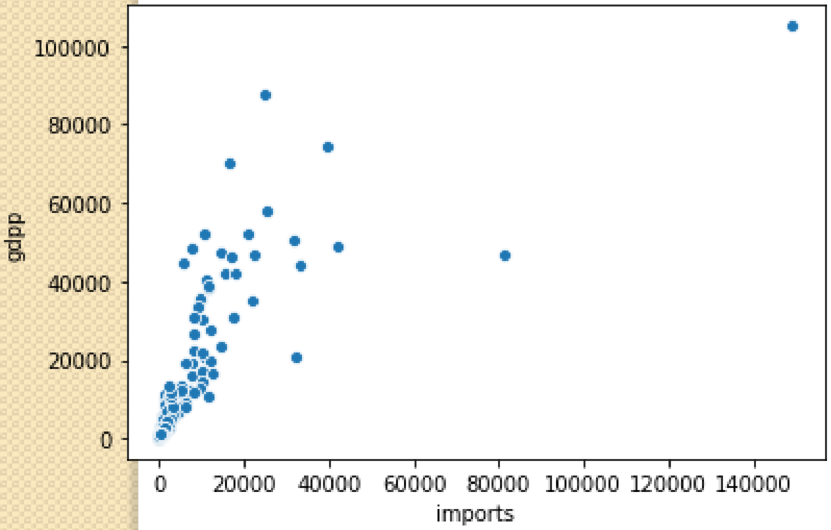


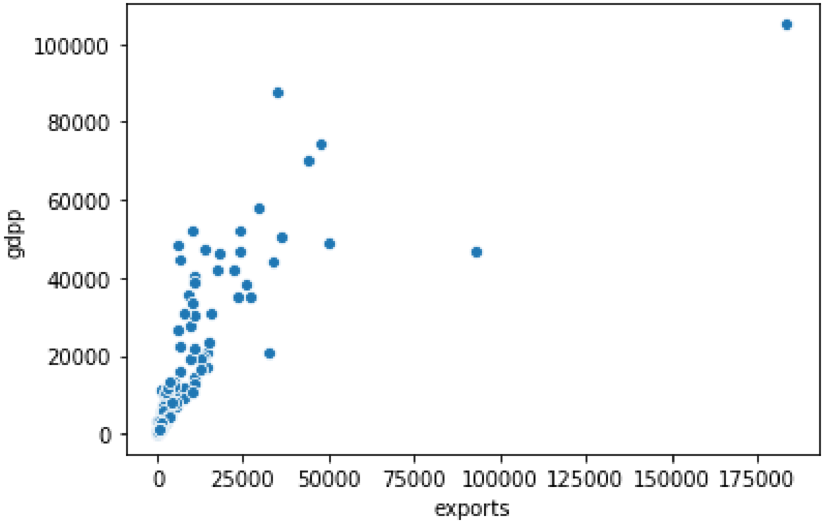




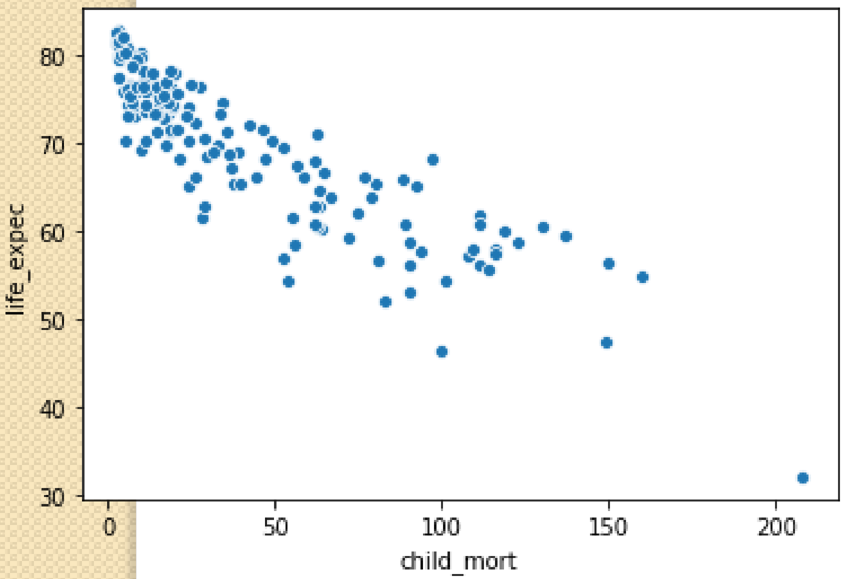
- *exports & income, exports & gdpp, imports & gdpp:* As exports/imports are increasing in small amount, the income and gdpp are increasing in large amount.

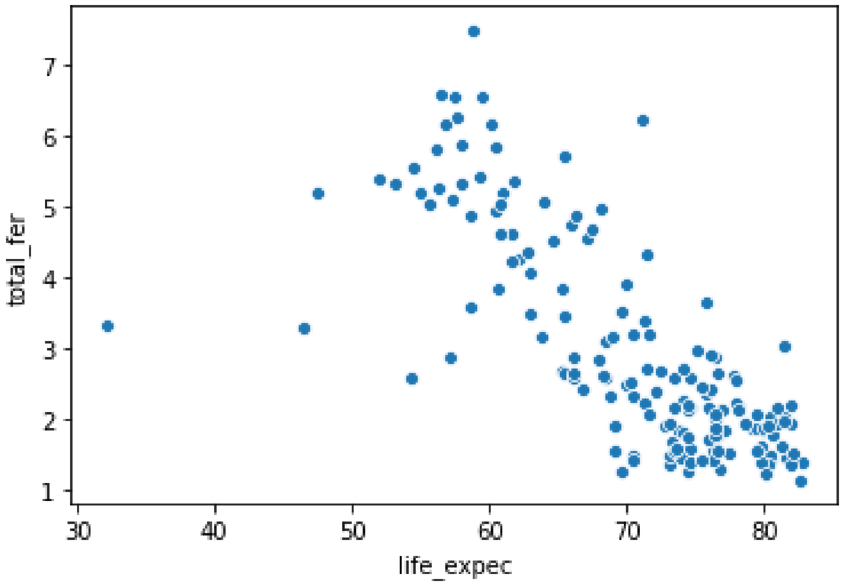




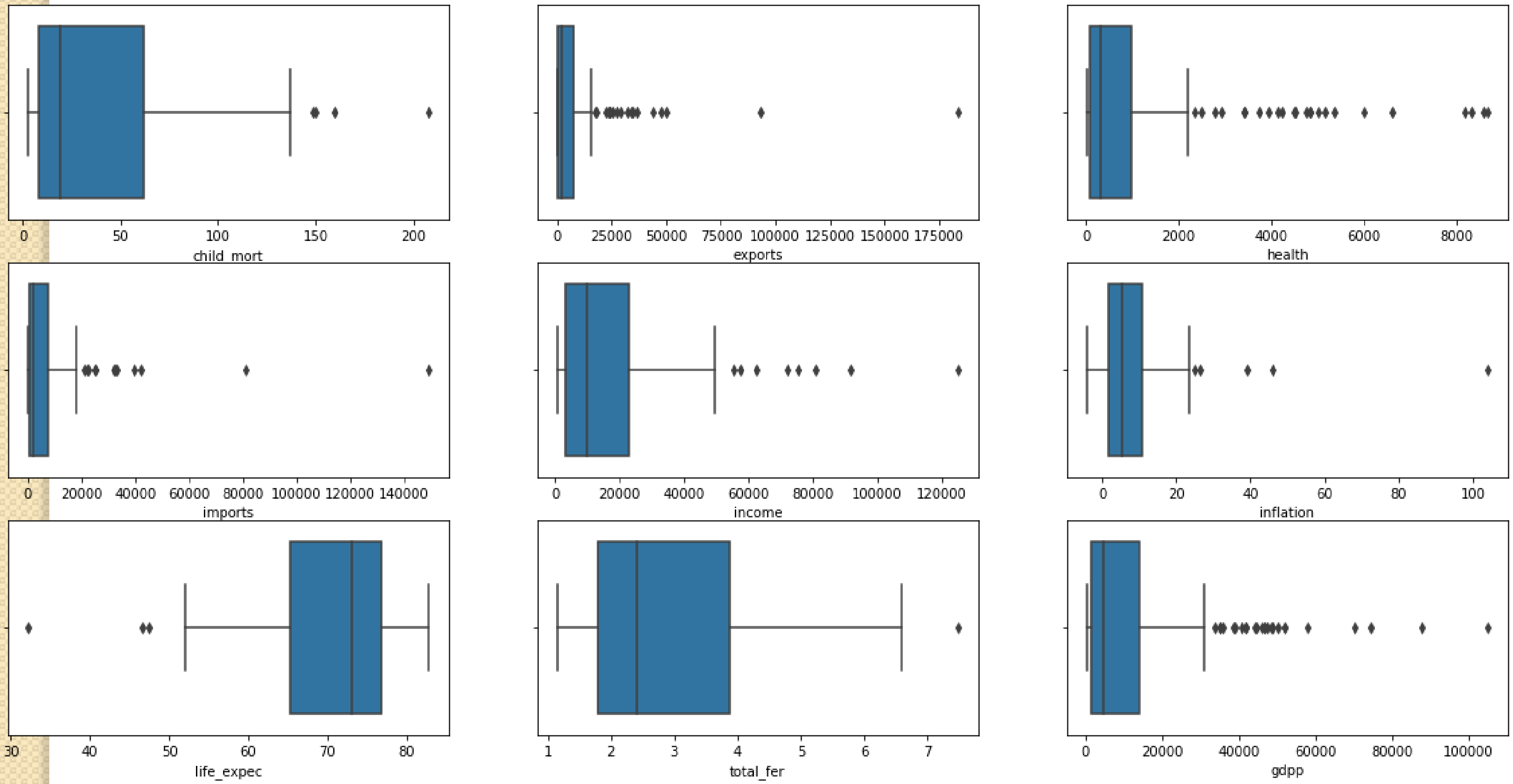


- **Negative Correlation:**
  - *child_mort & life_expec*: As the child mortality is increasing the life expectancy of infants decreasing drastically.
  - *life_expec & total_fer*: As life expectancy of an infant is more, the infants born are less.

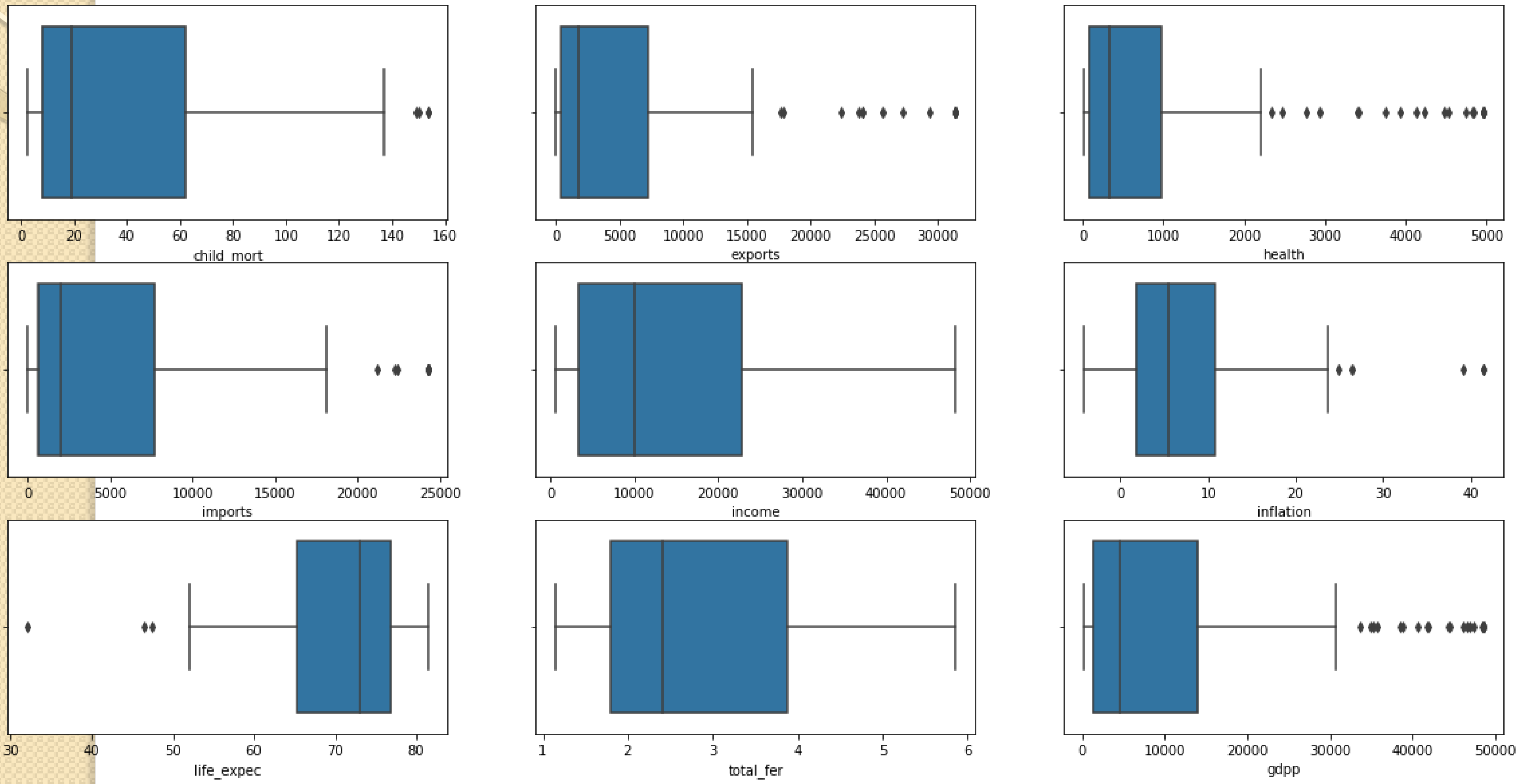




- **Handling Outliers:**
  - The box plot shows the outliers at the higher fence except for the column 'life_expec', which is having the outlier at the lower fence.
  - This outlier of life_expec at the lower fence will not be treated as it might be important for analysis.



- To over come outliers, data are capped instead of deleting as the there are less data in the data set.
- For variables 'child_mort' and 'inflation' soft-range capping (1-99) is used as a single data point is way out of range.
- For remaining variables, mid-range capping (5-95) was the better option.
- Outliers after data being capped are as shown in the next slide.
- After employing the capping method, some of the data points which were way out of range are within range.
- But yet some of the data points are at the upper fence and remaining analysis will be carried without further capping of outliers.


0
20
40
60
80
100
120
140
160
child_mort
0
5000
10000
15000
20000
25000
30000
exports
0
1000
2000
3000
4000
5000
health
0
5000
10000
15000
20000
25000
imports
0
10000
20000
30000
40000
50000
income
0
10
20
30
40
inflation
30
40
50
60
70
80
life_expec
1
2
3
4
5
6
total_fer
0
10000
20000
30000
40000
50000
gdpp

Clustering Model

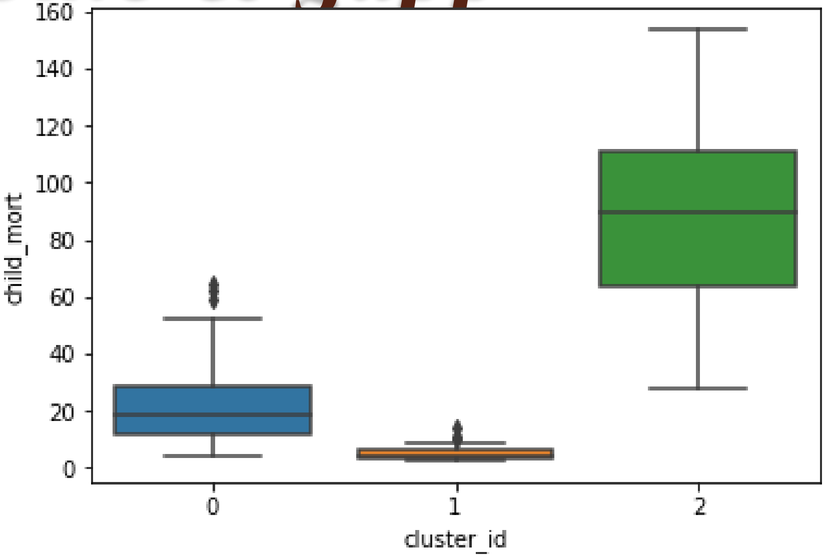
- The hopkin’s test was conducted to check the cluster tendency.
- Scaling of data using Standard Scaler so that all data are in same range.
- **K-Means Clustering Algorithm:**
  - To find the optimal number of clusters two methods were employed: ***Sihouette Score*** *and* ***Elbow-Curve***.
  - The number of clusters was decided as **3** and the model was fitted.
  - The clusters formed were as follows and they were analyzed with respect to the variables [‘child_mort’, ‘income’, ‘gdpp’] .
    1. C1 → high child mortality, low income and gdpp.
    2. C2 → low child mortality, high income and gdpp.
    3. C3 → slightly better than C1.

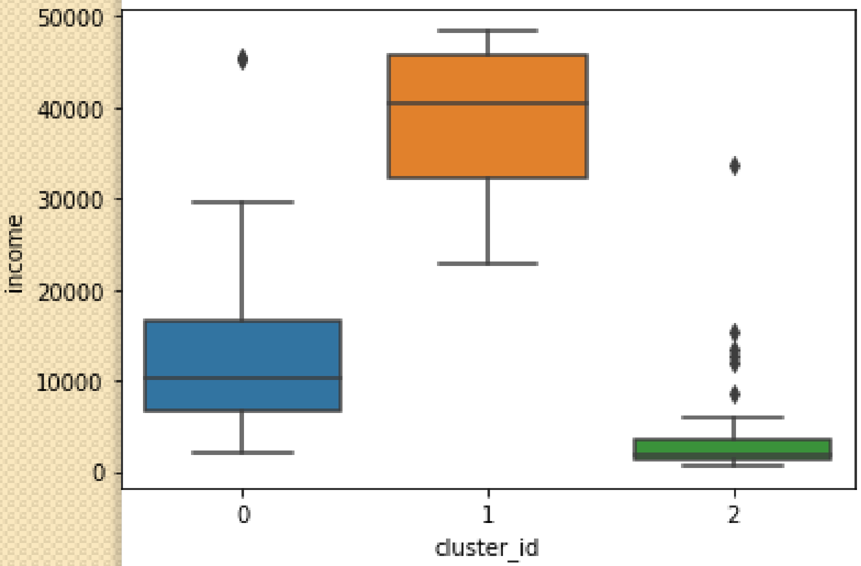
Visualizing clusters with respect to child_mort, income & gdpp

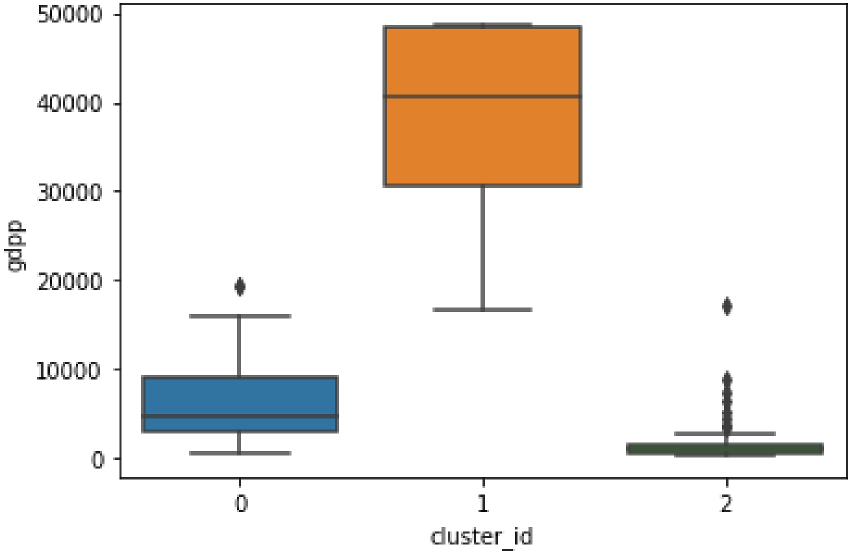
Cluster 2: Includes countries with high child mortality, low income & low gdpp.

Cluster 1: Includes countries with low child mortality, high income & high gdpp.

Cluster 0: Includes countries with child mortality slightly more than cluster 1, income and gdpp slightly more than cluster 2.





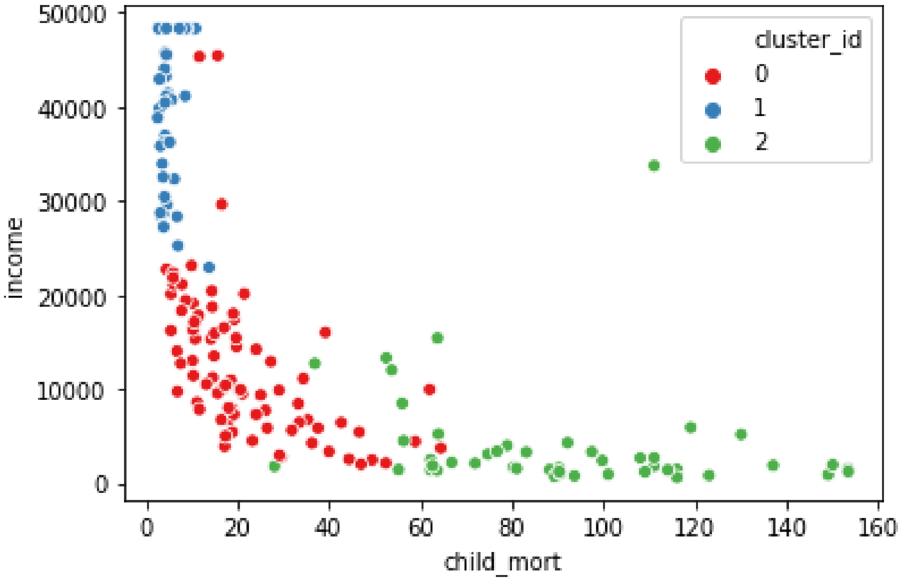


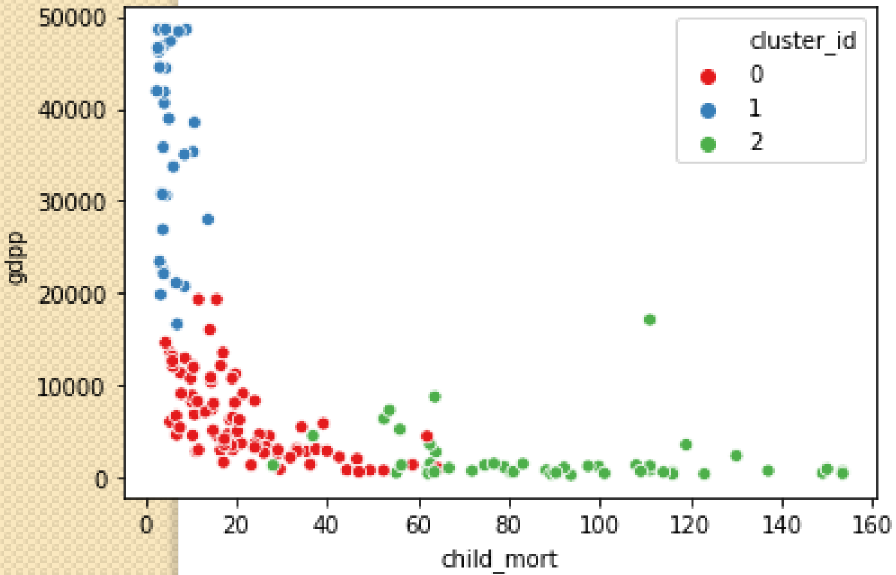
Relationship between the variables, child_mort, income & gdpp.

Cluster 1: high 'income' but low 'child_mort',

Cluster 2: low 'income' but high 'child_mort' and

Cluster 0: slightly higher 'income' and lower 'child_mort' when compared to cluster 2.





Cluster 1: high 'gdpp' but low 'child_mort',

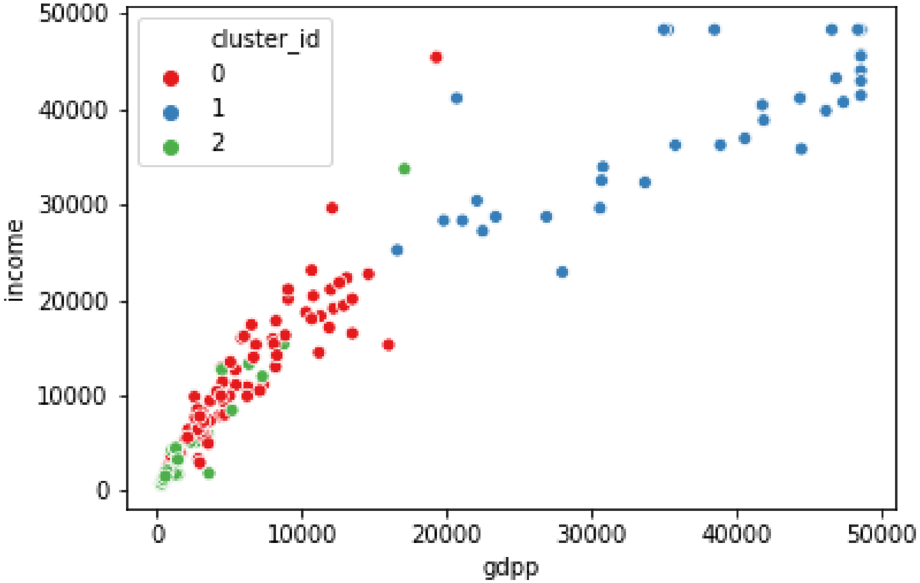
Cluster 2: low 'gdpp' but high 'child_mort' and

Cluster 0: slightly higher 'gdpp' & lower 'child_mort' compared to cluster 2.

Cluster 1: high 'gdpp'and 'income‘.

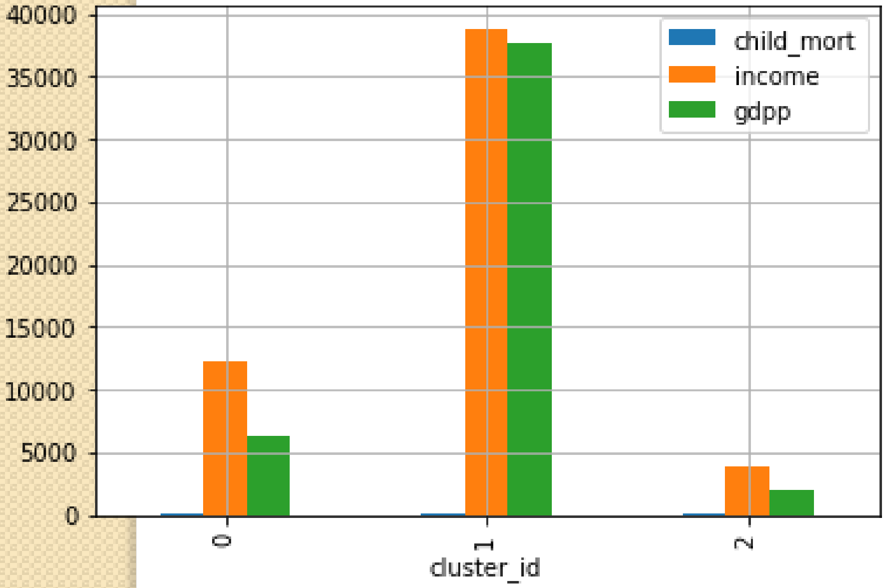
Cluster 2: low 'gdpp'and 'income'.

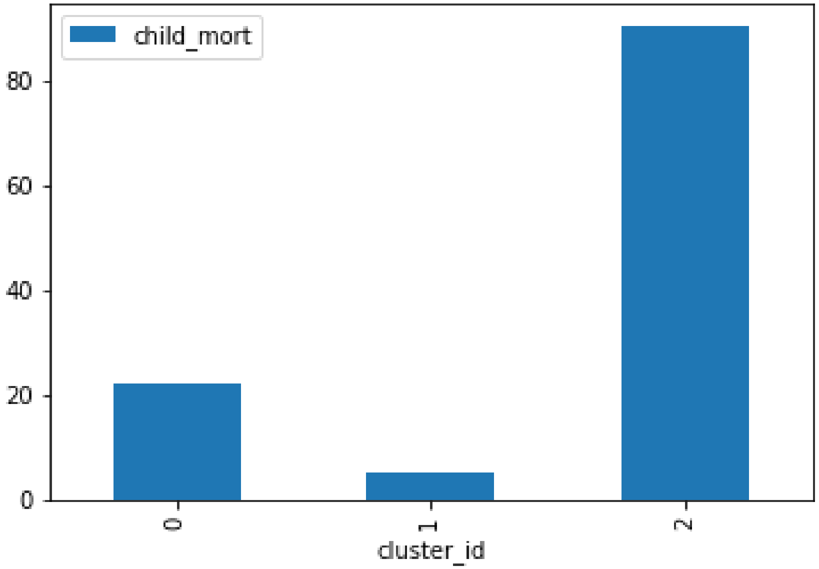
Cluster 0: slightly higher 'gdpp' and 'income' when compared to cluster 2.



Cluster Profiling:

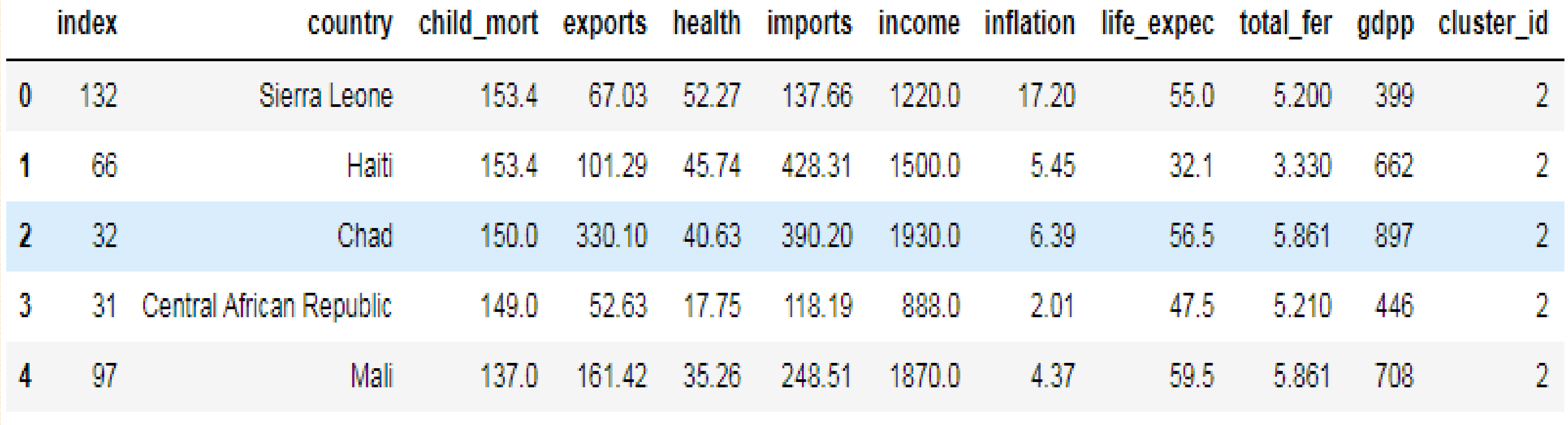
From the below two bar plots it can be clearly seen that cluster 2 is having low 'income' and 'gdpp and high 'child_mort' when compared to other two clusters.





Results:

Top 5 countries which need immediate aid are listed below:



	index	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	cluster_id
0	132	Sierra Leone	153.4	67.03	52.27	137.66	1220.0	17.20	55.0	5.200	399	2
1	66	Haiti	153.4	101.29	45.74	428.31	1500.0	5.45	32.1	3.330	662	2
2	32	Chad	150.0	330.10	40.63	390.20	1930.0	6.39	56.5	5.861	897	2
3	31	Central African Republic	149.0	52.63	17.75	118.19	888.0	2.01	47.5	5.210	446	2
4	97	Mali	137.0	161.42	35.26	248.51	1870.0	4.37	59.5	5.861	708	2

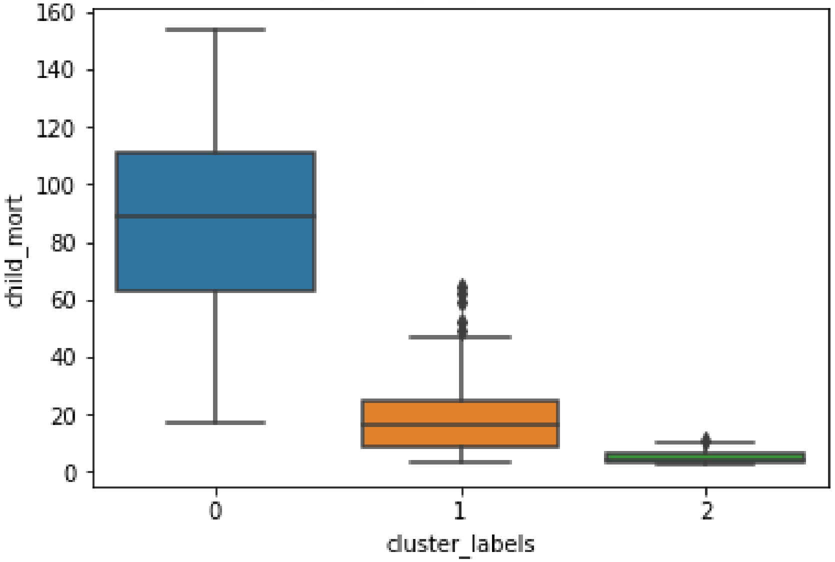
- **Hierarchical Clustering Algorithm:**
  - Dendrogram was created using complete linkage instead of single linkage as single linkage did not create clear dendrogram.
  - The number of clusters was decided to be 3 though from the dendrogram it looks like 4.
  - As the clusters formed for 3 was better than the clusters formed for 4.

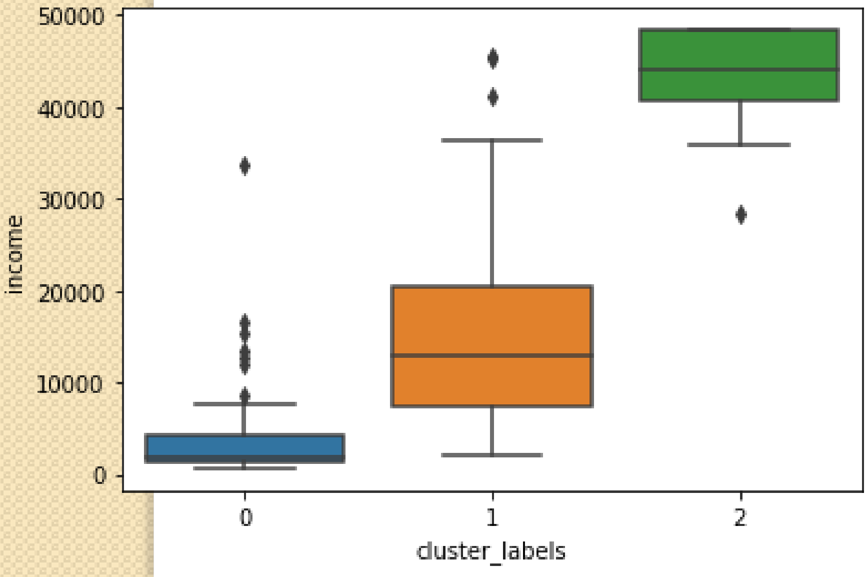
Visualizing clusters with respect to child_mort, income & gdpp

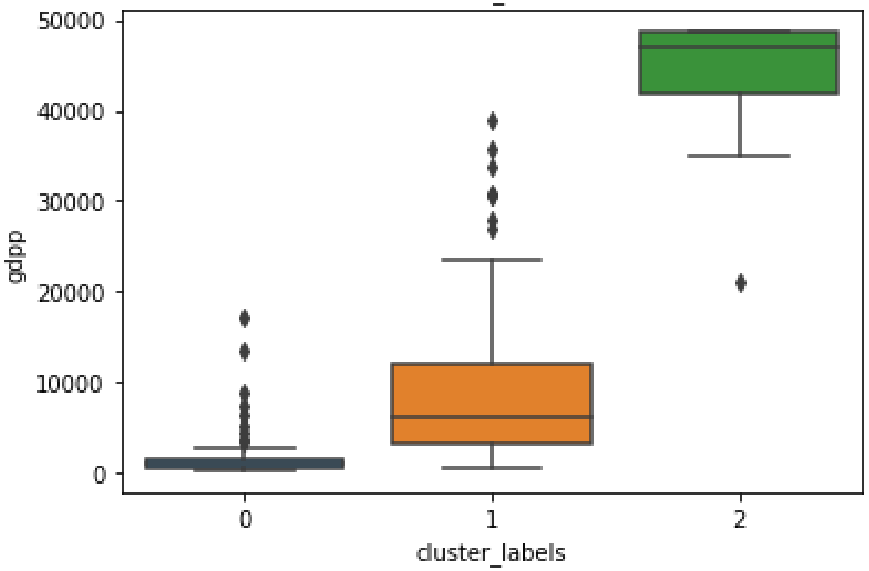
Cluster 0: Includes countries with high child mortality, low income and low gdp.

Cluster 2: Includes countries with low child mortality, high income and high gdp.

Cluster 1: Includes countries with child mortality slightly more than cluster 2, income and gdp is more than cluster 0.





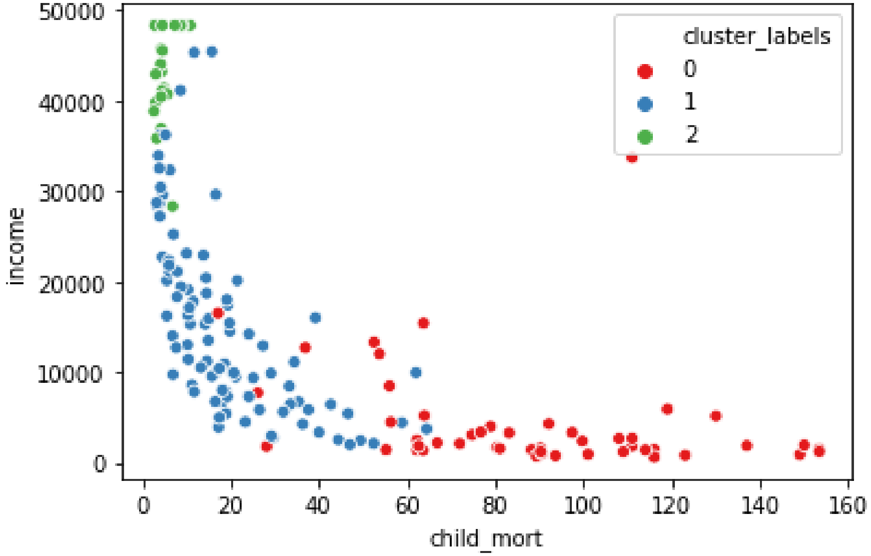


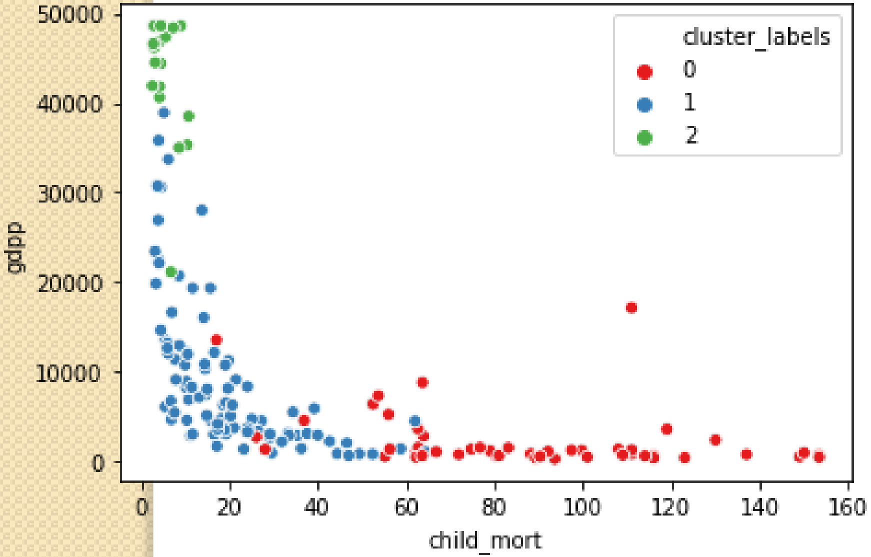
Relationship between the variables, child_mort, income & gdpp

Cluster 2: high ‘income’ but low ‘child_mort’,

Cluster 0: low ‘income’ but high ‘child_mort’ and

Cluster 1: slightly higher ‘income’ and lower ‘child_mort’ when compared to cluster 0.





Cluster 2: high ‘gdpp’ but low 'child_mort’,

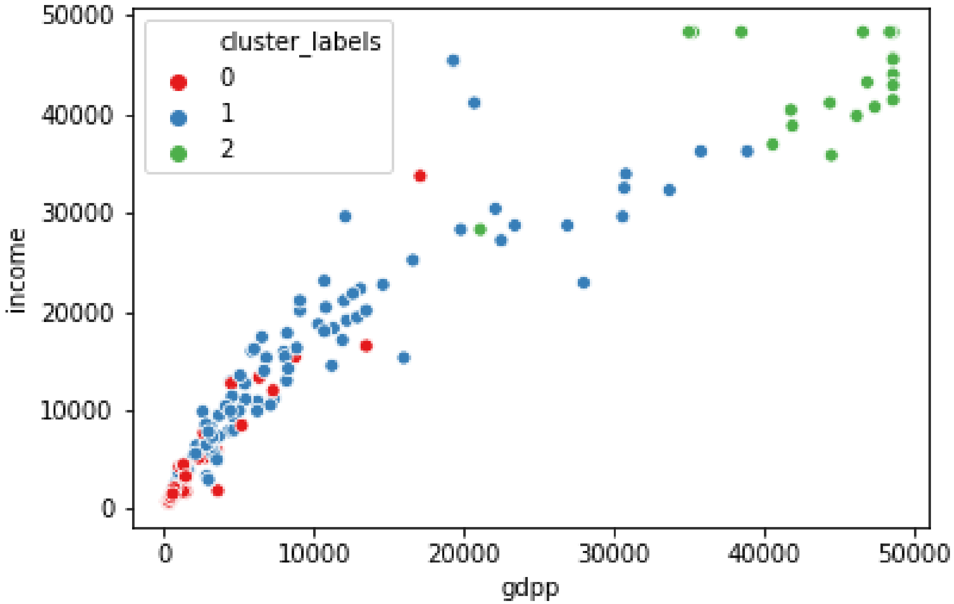
Cluster 0: low ‘gdpp’ but high 'child_mort’ and

Cluster 1: slightly higher ‘gdpp’ & lower ‘child_mort’ compared to cluster 0.

Cluster 2: high 'gdpp'and 'income‘.

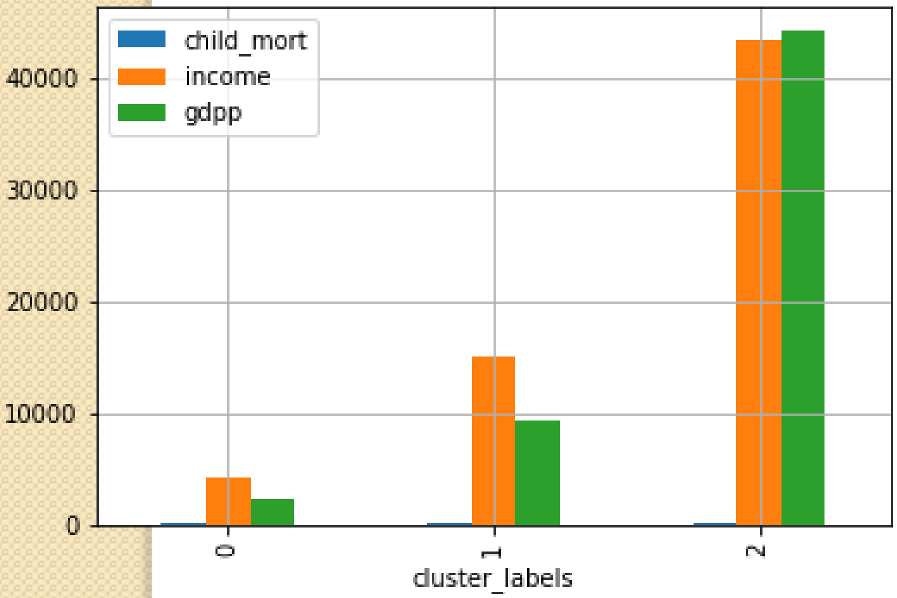
Cluster 0: low 'gdpp'and 'income'.

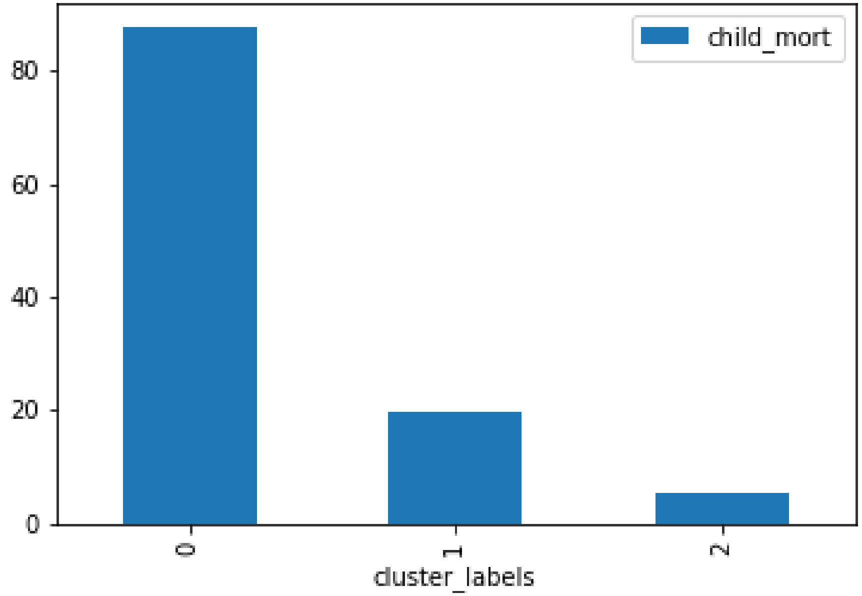
Cluster 1: slightly higher 'gdpp' and 'income' when compared to cluster 2.



Cluster Profiling:

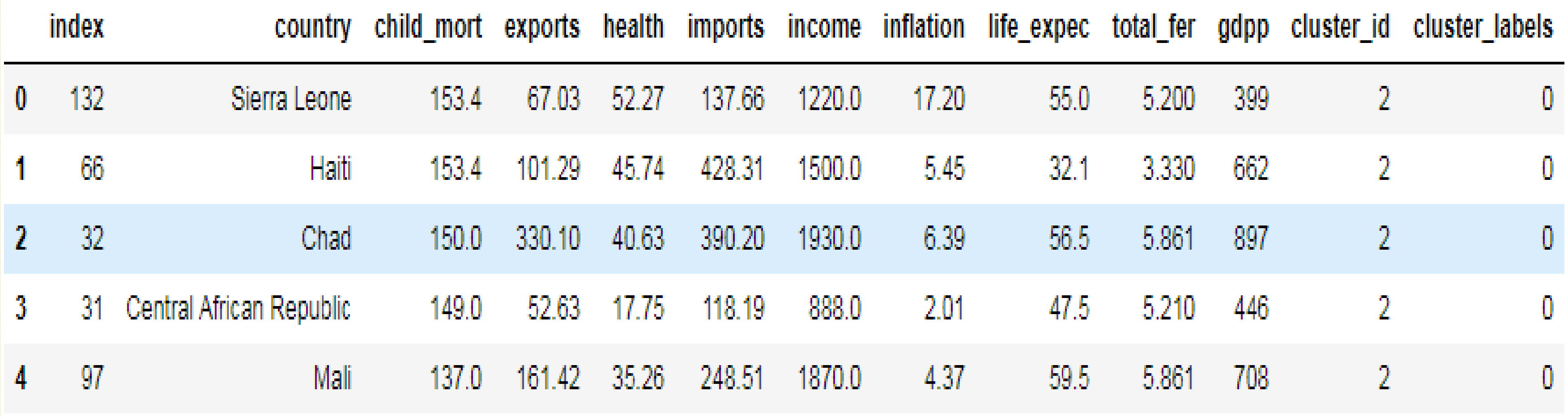
From the above two bar plots it can be clearly seen that cluster 0 is having low 'income' and 'gdpp and high 'child_mort' when compared to other two clusters.





Results:

Top 5 countries which need immediate aid are listed below:



	index	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	cluster_id	cluster_labels
0	132	Sierra Leone	153.4	67.03	52.27	137.66	1220.0	17.20	55.0	5.200	399	2	0
1	66	Haiti	153.4	101.29	45.74	428.31	1500.0	5.45	32.1	3.330	662	2	0
2	32	Chad	150.0	330.10	40.63	390.20	1930.0	6.39	56.5	5.861	897	2	0
3	31	Central African Republic	149.0	52.63	17.75	118.19	888.0	2.01	47.5	5.210	446	2	0
4	97	Mali	137.0	161.42	35.26	248.51	1870.0	4.37	59.5	5.861	708	2	0

Conclusion

- The 3 clusters formed in both clustering algorithm are:
  1. C1 - High Child Mortality, Low Income & GDPP.
  2. C2 - Low Child Mortality, High Income & GDPP.
  3. C3 - Better cluster when compared to C1.
- The segmentation of data was better in K-Means when compared to Hierarchical for number of clusters = 3.
- Both K-Means and Hierarchical Clustering Algorithms are producing the same results for number of clusters = 3.
- Hence, the top-5 countries which need immediate aid are:
  1. Sierra Leone (child_mort = 153.4, income = 1220, gdpp= 399)
  2. Haiti (child_mort = 153.4, income = 1500, gdpp= 662)
  3. Chad (child_mort = 150.0, income = 1930.0, gdpp= 897)
  4. Central African Republic (child_mort=149.0,income=888,gdpp=446)
  5. Mali (child_mort = 137.0, income = 1870.0, gdpp= 708)

THANK YOU