Clustering Assignment

Submitted by :
Karan Sehgal

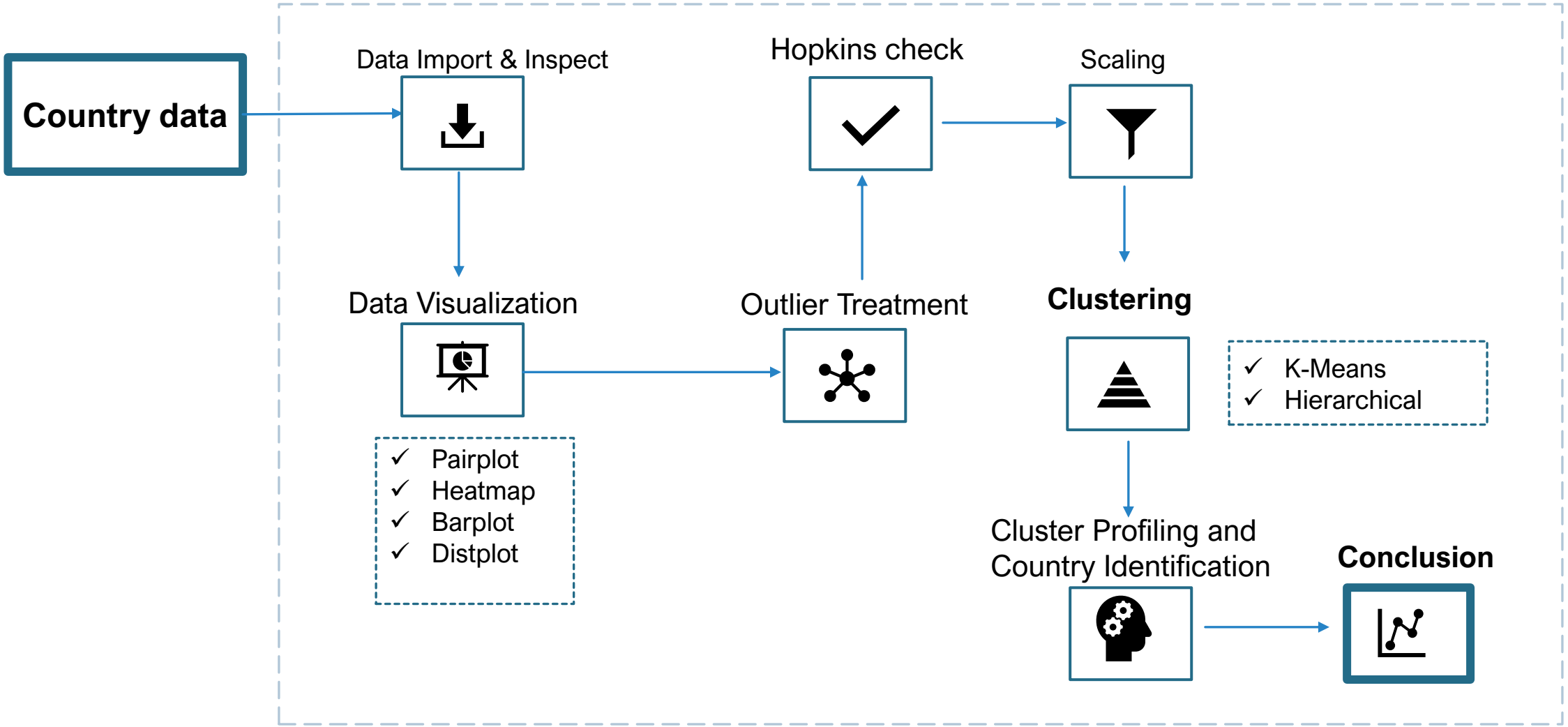
Business Understanding

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.

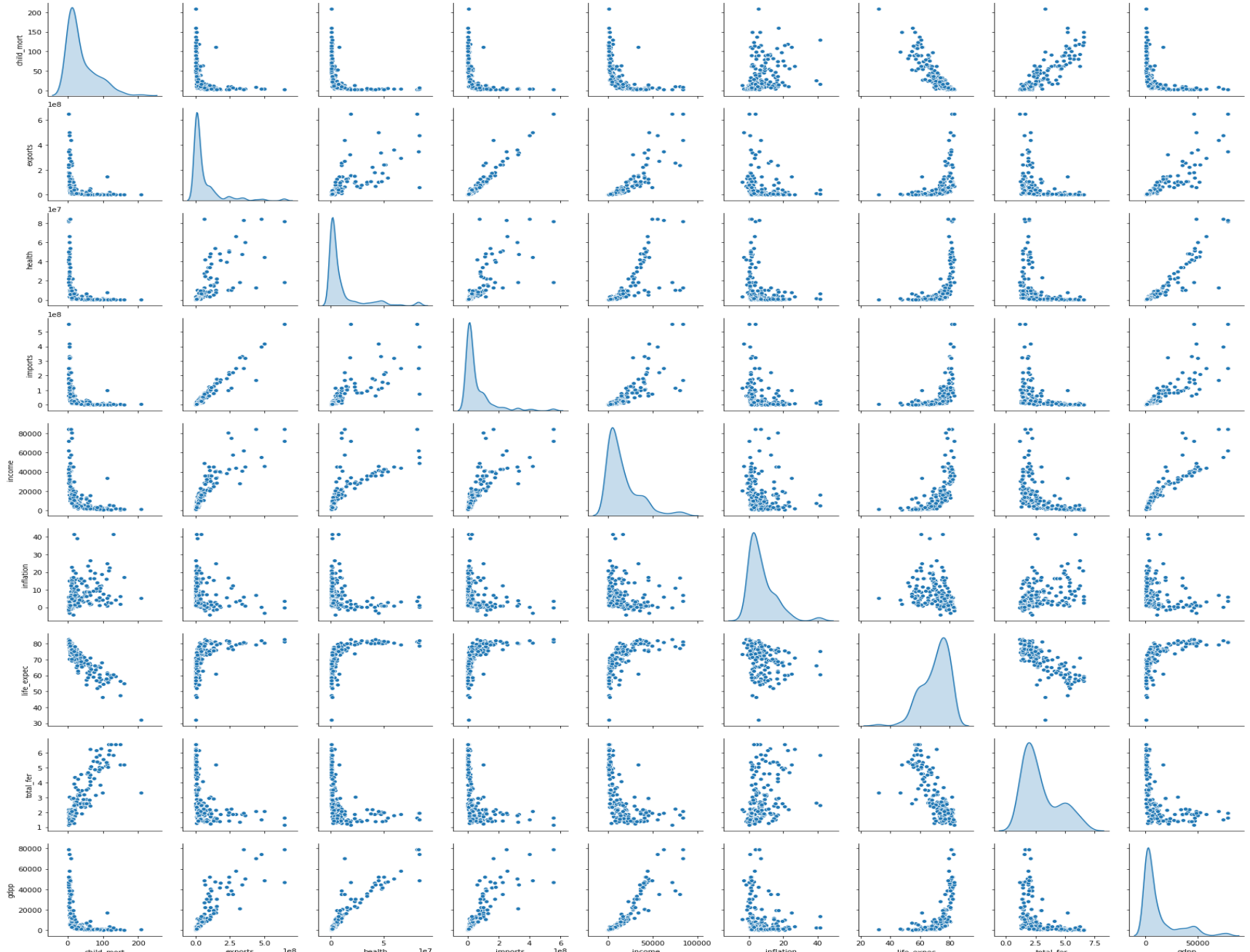
After the recent funding programs, they have been able to raise around $ 10 million. Now the CEO of the NGO needs to decide how to use this money strategically and effectively. The significant issues that come while making this decision are mostly related to choosing the countries that are in the direst need of aid.

As an analyst, we have to come up with the countries list that are in the direst need of aid.

Data Workflow



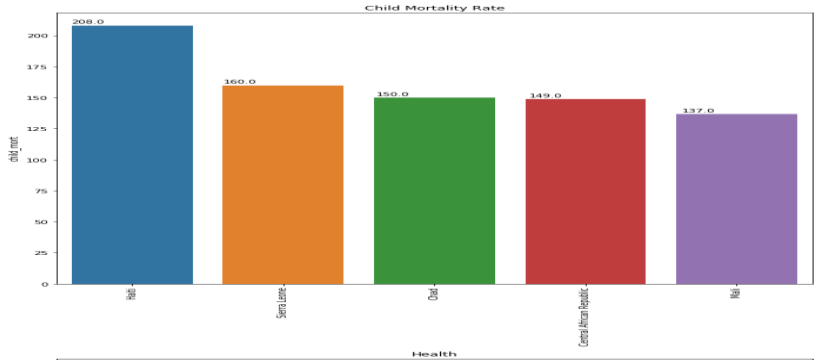
Data Visualisation

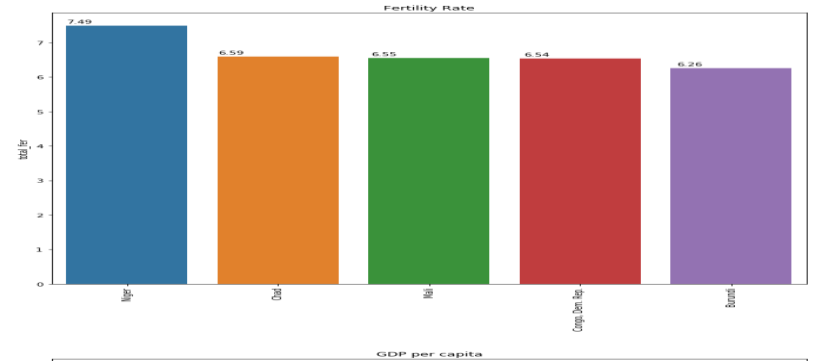


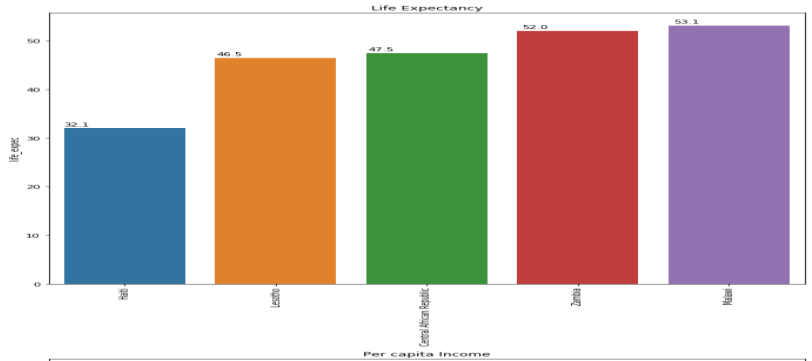
Pair plot between numerical variables:

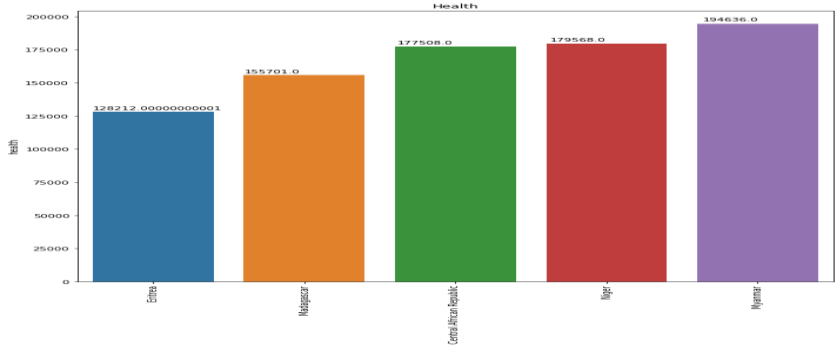
- Linear relation is found between gdpp-income, imports-exports, total_fer-child_mort
- Rectangular hyperbola curve is generated by gdpp-child_mort.
- If gdpp is HIGH:
  - child mortality is LOW
  - income is HIGH
  - inflation is LOW
  - life expectancy is HIGH
  - total fertility is LOW
  - health, imports and exports are MEDIUM

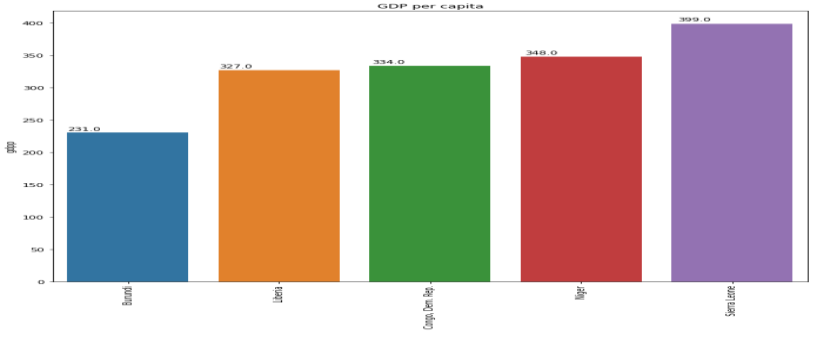
Data Visualisation

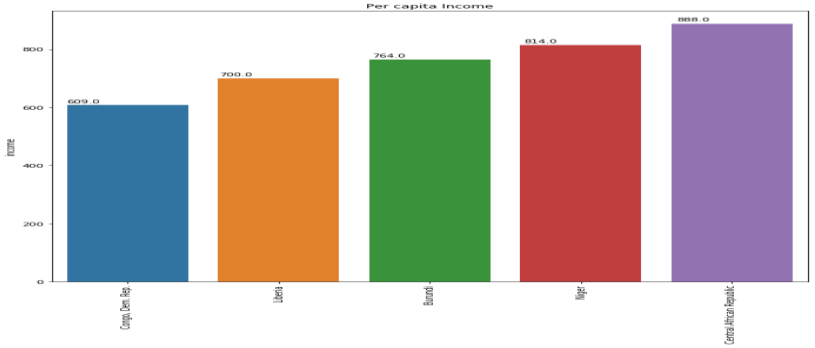


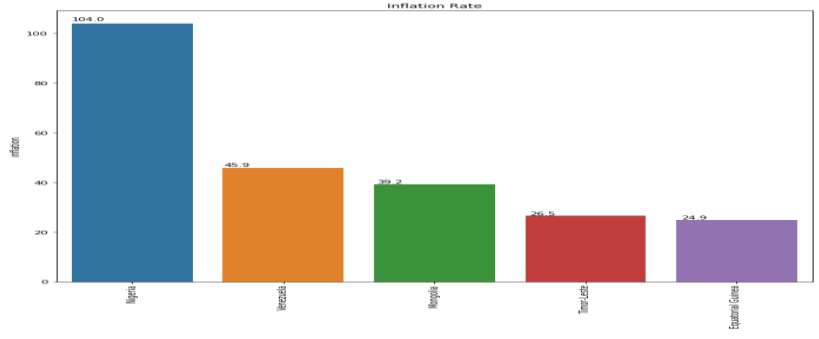


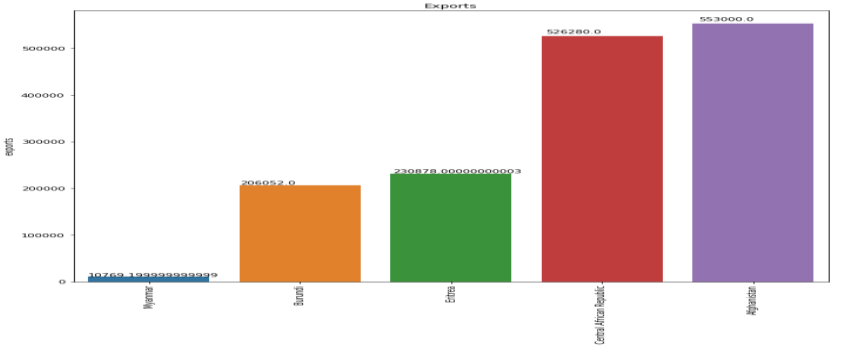


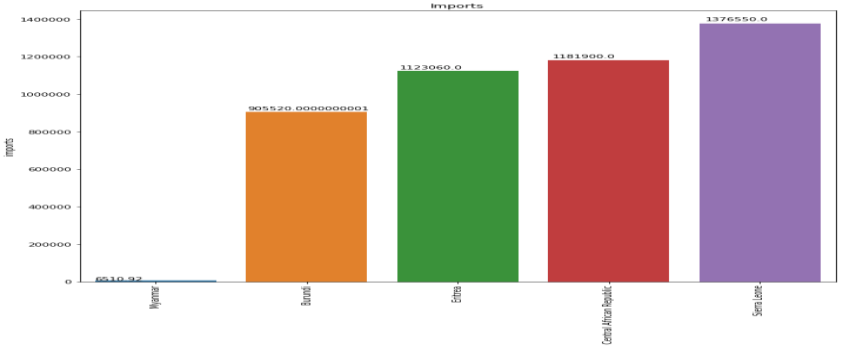








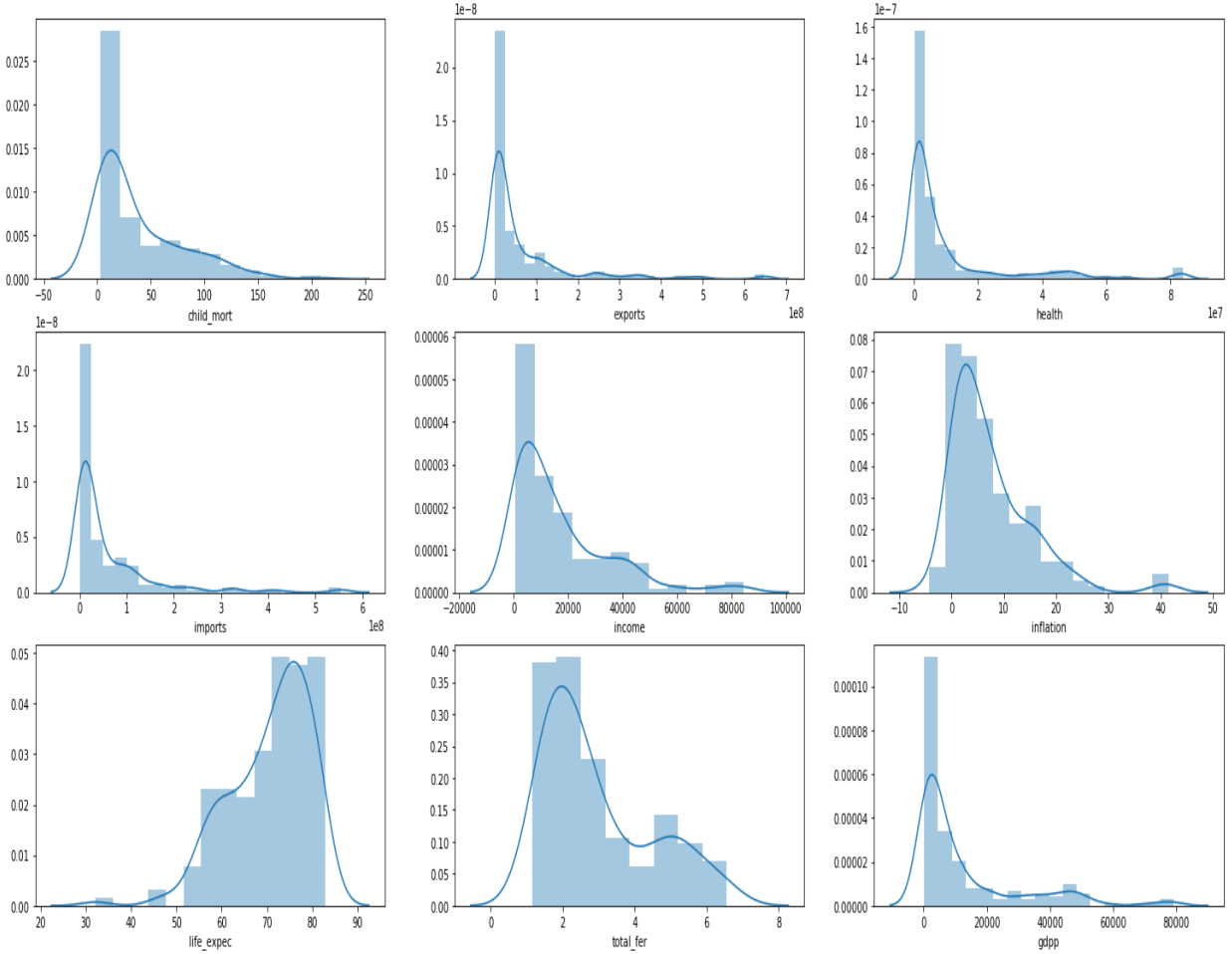


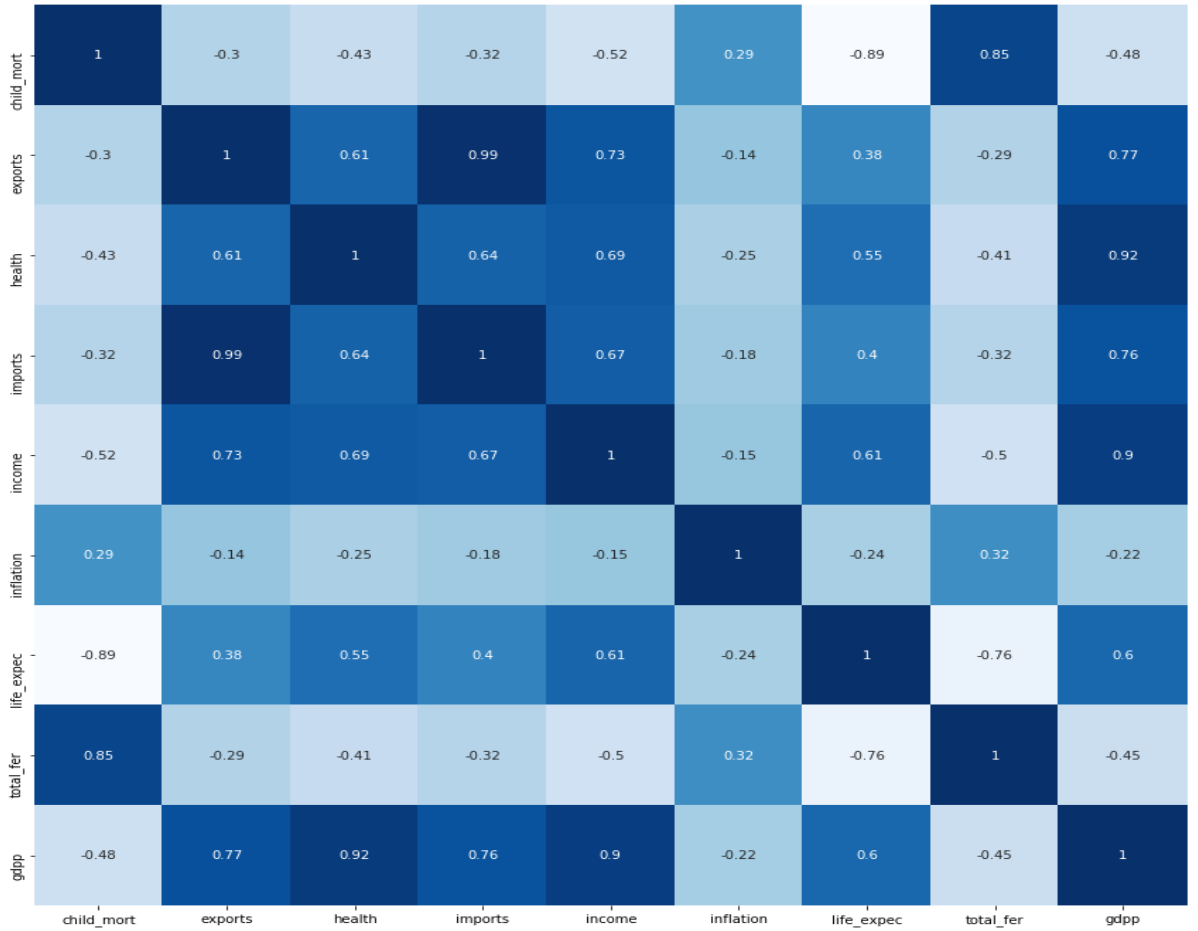


Barplot for every variable :

- The above plots shows the five countries which are in need of aid individually for all the factors taken in consideration.
- Based on these graphs, I decided to choose GDPP, Income and Child Mortality as the driving socio-economic factor.
- After clustering, we'll be choosing the countries as a combination of above factor.

Data Visualisation





Distribution Plot:

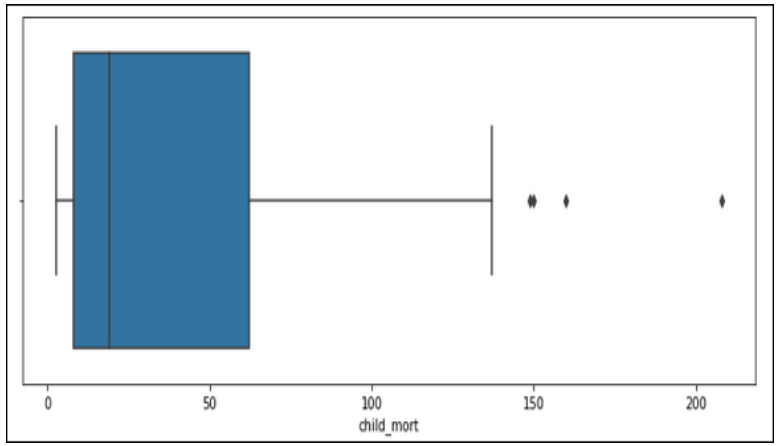
- life_expec is right-skewed whereas all the rest features are left-skewed.
- total_fer and gdpp are bimodal whereas all the rest features are unimodal.

Heatmap:

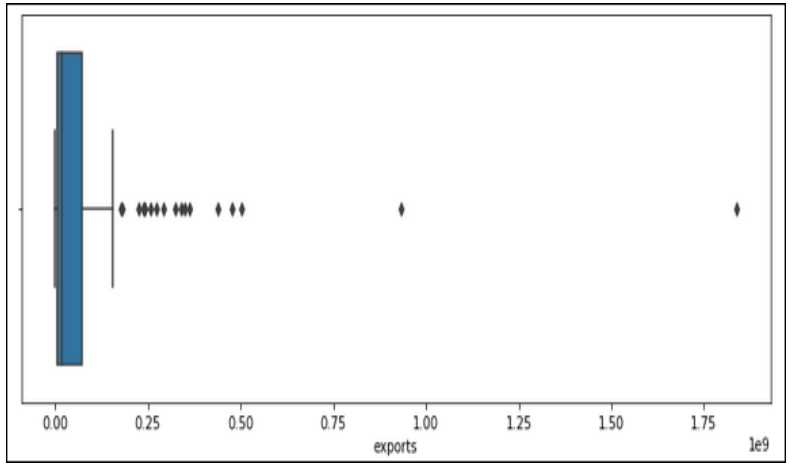
- exports is highly correlated with imports.
- health, exports, income,imports are highly correlated with gdpp.
- child_mort is having high negative correlation with life_expec.
- total_fer is highly positively correlated with child_mort and negatively correlated with life_expec

Outlier Treatment

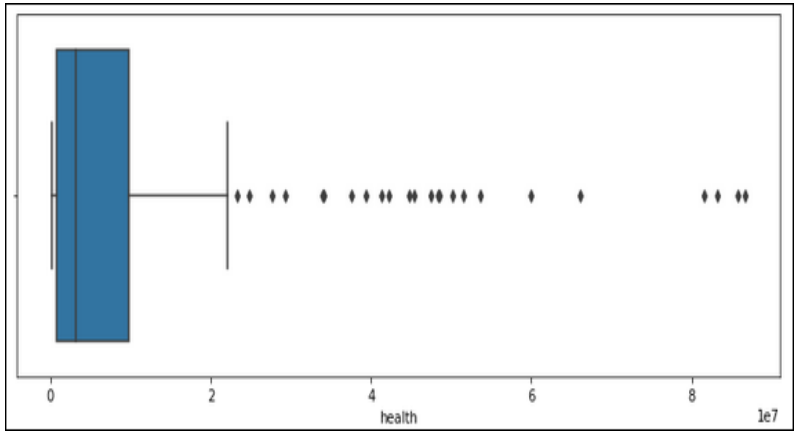
Child_mort: There are Upper end outliers for this feature but capping is not done as those countries would be in need of aid.



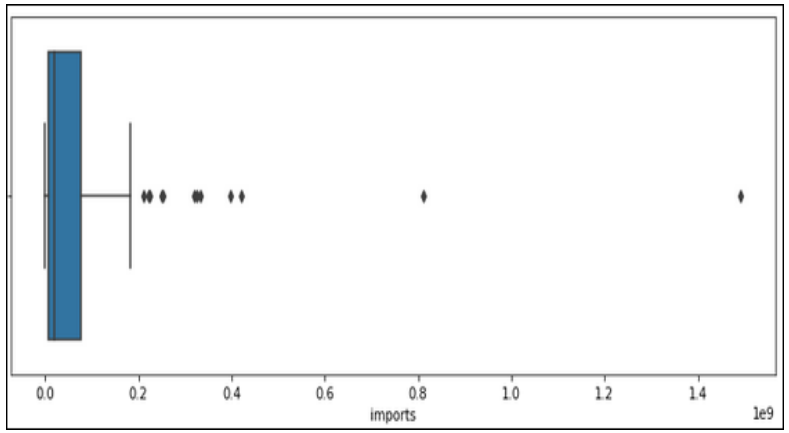
Exports : There are a lot of outliers for the feature and upper end outliers were capped to 99th percentile as our business focus is on the under-developed countries.



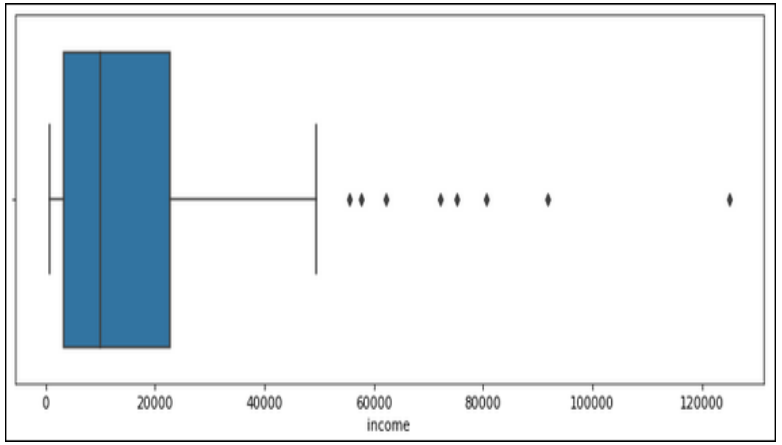
Health : For total spending on health per capita, the countries at upper end were reported as outliers and were capped under soft range capping



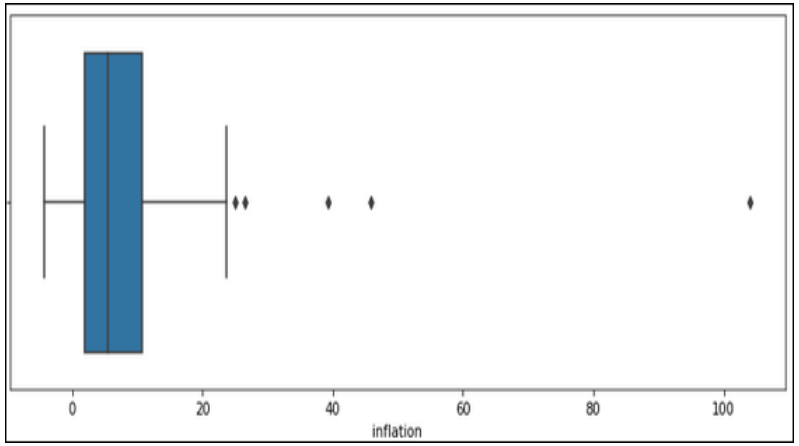
Imports : Imports of goods and services per capita, the countries at upper end emerged as outliers and were capped to 99th percentile as per business requirements.



Income: For net income per person, the countries at upper end were capped to 99th percentile as per business need.

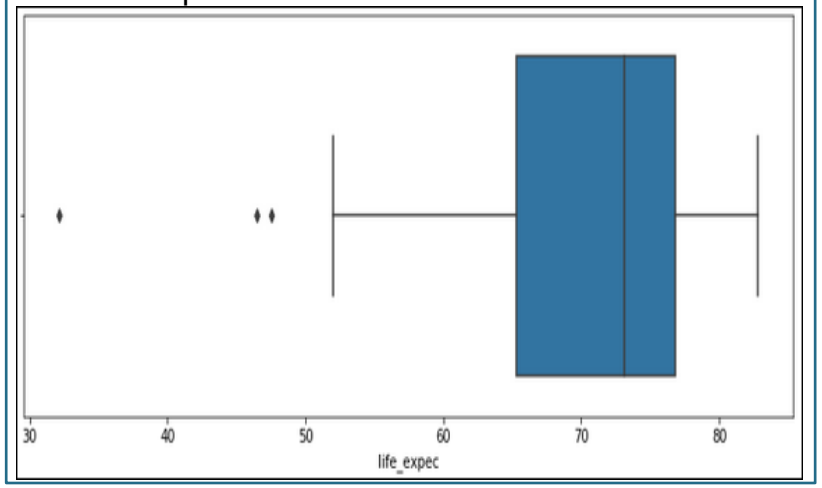


Inflation: The Measurement of average growth doesn't have a lot of outliers, however, the upper end outliers were capped.

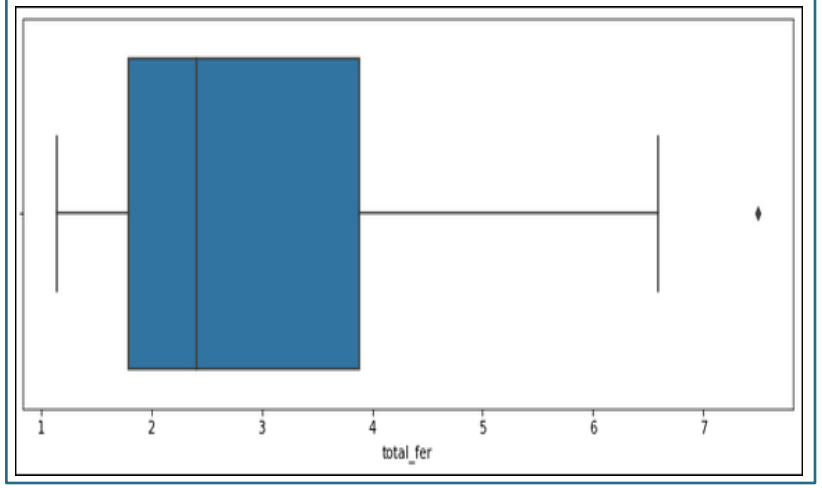


Outlier Treatment

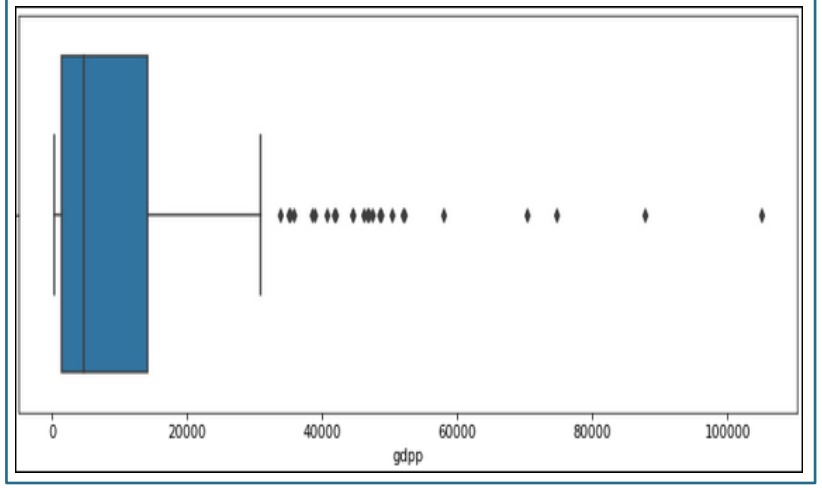
Life_expec: The average number of years a new born child would live, has few outliers ate lower end, which are not capped as these countries are of much importance to us.



Total_fer : The number of children that would be born to each woman, has only one outlier, which is capped as per business need.



GDPP : The Total GDP divided by the total population, has outliers for the countries having higher value, where soft capping was performed.

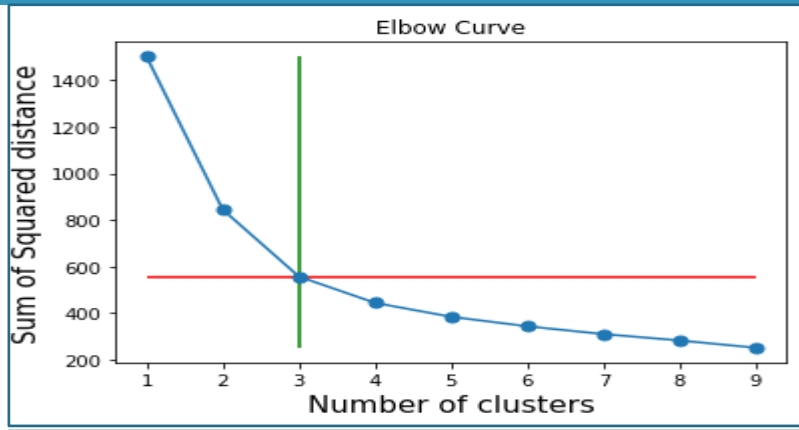


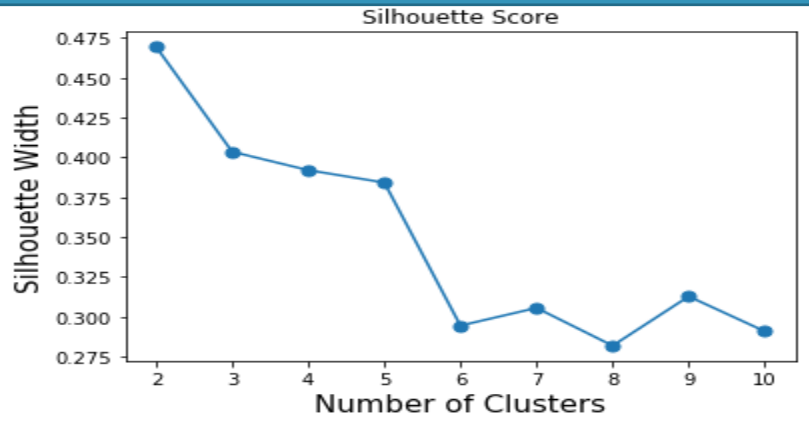
Hopkins check and Scaling

Hopkins Statistics : To check clustering tendency, I calculated the hopkins statistics. It determines whether the data points differs significantly from uniformly distributed data in multidimensional space. A score of 90+ depicted that the dataset has good clustering tendency.

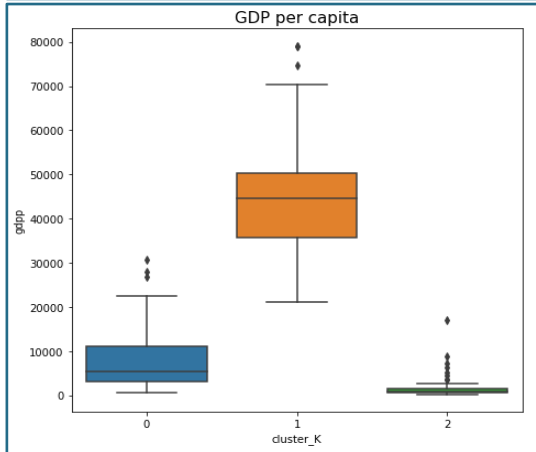
Scaling : Data Normalization standardize the raw data by converting them into specific range using a linear transformation which can generate good quality clusters and improve the accuracy of clustering algorithms. Using Standard-Scaler, the features were scaled around the centre with mean 0 and with a standard deviation of 1. After this, the data was ready for clustering.

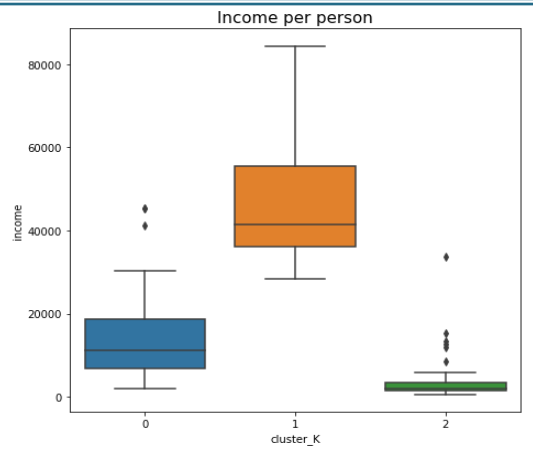
Clustering : K-Means

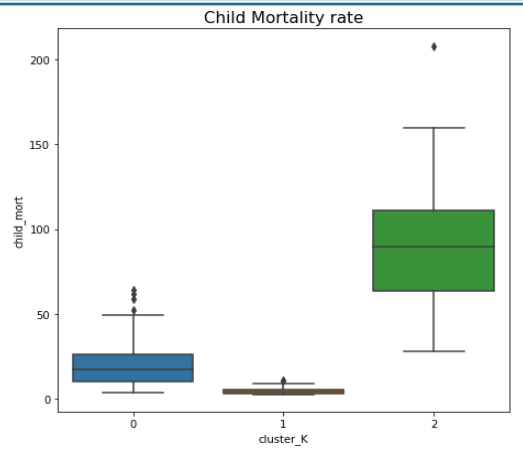


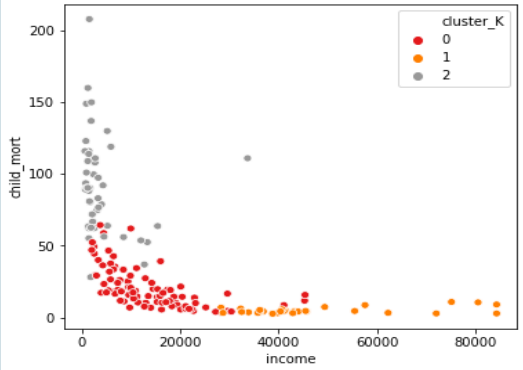


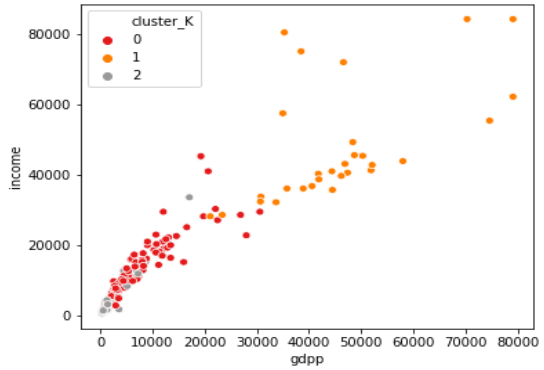
Based on Elbow curve and Silhouette Analysis Curve, I decide to make a trade-off and choose the optimum value of K as 3

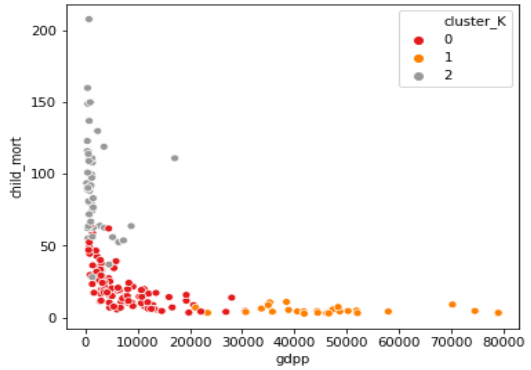












After performing the clustering using K-means, the clusters were formed and for visualising, I chose scatterplot and boxplot for variables : Income, GDPP and Child Mortality.

- Countries with low gdpp, income and high child mortality are Under-developed countries (cluster_K = 2)
- Countries with high gdpp, income and low child mortality are Developed countries (cluster_K = 1)
- Countries with low gdpp, income and low child mortality are Developing countries (cluster_K = 0)

Clustering Profiling and Country Identification

index	cluster_K
0	90
2	48
1	29

Table 1

cluster_K	child_mort	income	gdpp
0	20.55	13804.33	7808.58
1	4.98	47784.41	46068.14
2	91.61	3897.35	1909.21

Table 2

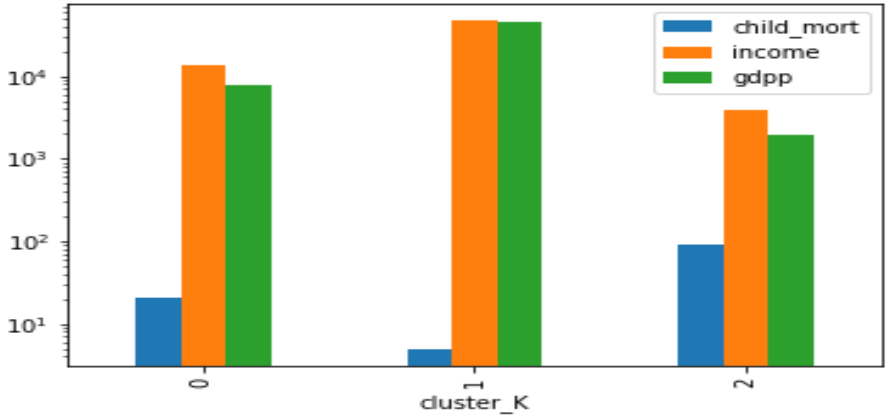
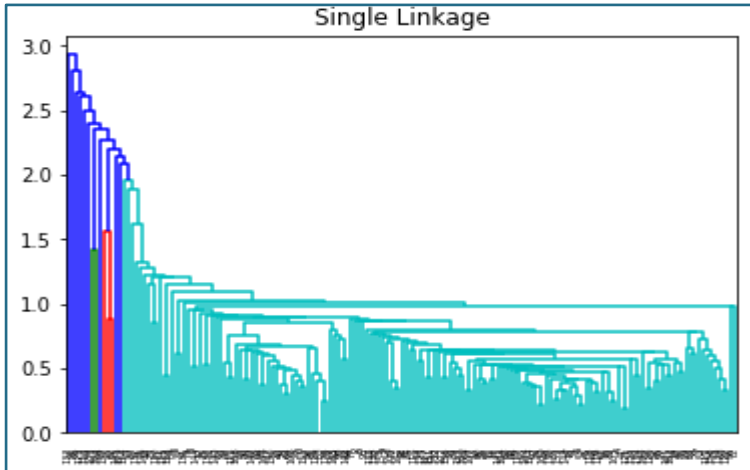


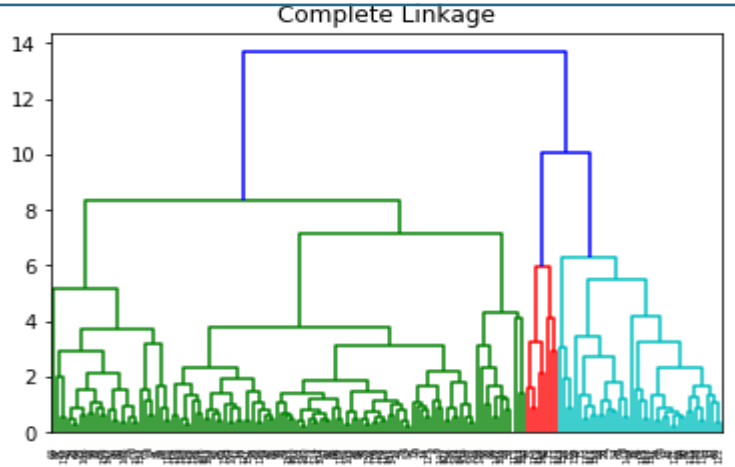
Table 1 represents the number of countries in each cluster and Table 2 represents the mean value of our main variable in each cluster.
Based on above plot and table 2, I know that countries with cluster index 2 are Under-Developed and requires aid.

Based on K-means clustering, The top 5 countries that need aid are listed below :

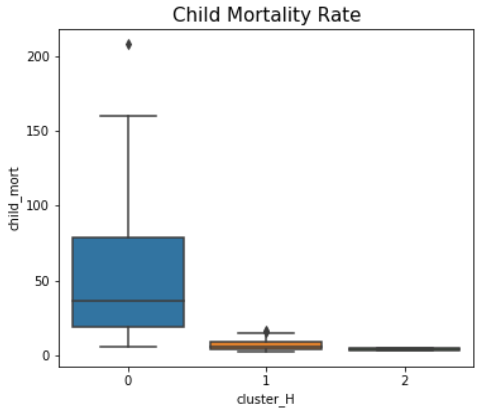
country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	cluster_K
Congo, Dem. Rep.	116.00	1372740.00	264194.00	1656640.00	609.00	20.80	57.50	6.54	334.00	2
Liberia	89.30	624570.00	385860.00	3028020.00	700.00	5.47	60.80	5.02	327.00	2
Burundi	93.60	206052.00	267960.00	905520.00	764.00	12.30	57.70	6.26	231.00	2
Niger	123.00	772560.00	179568.00	1708680.00	814.00	2.55	58.80	6.56	348.00	2
Central African Republic	149.00	526280.00	177508.00	1181900.00	888.00	2.01	47.50	5.21	446.00	2

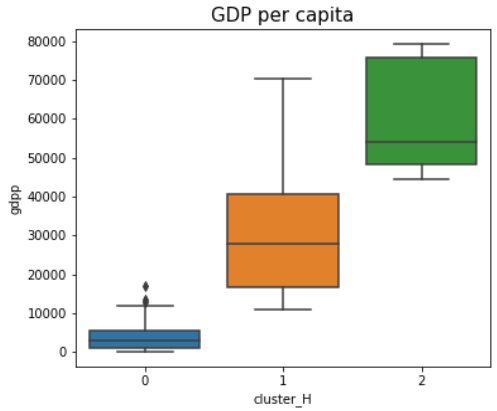
Clustering : Hierarchical

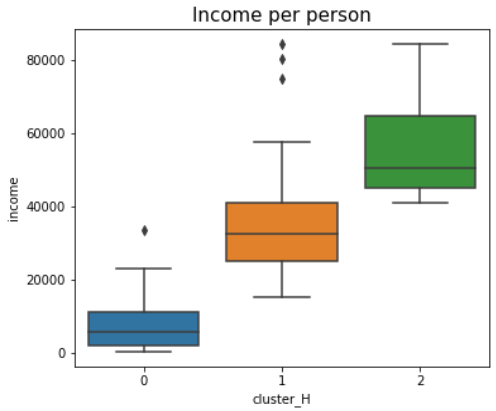


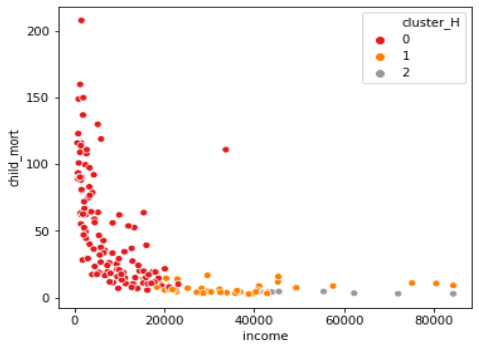


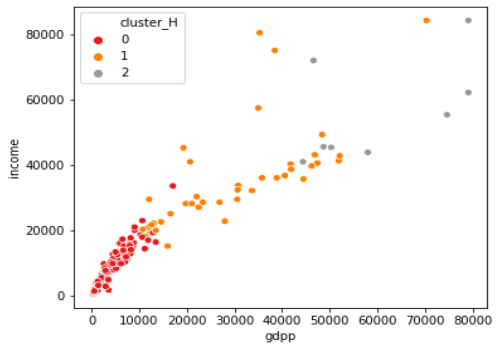
The dendrogram produced by single linkage is not well structured whereas The dendrogram produced by complete linkage is having proper tree-like structure.
Based on above complete linkage, Creating the hierarchical clustering model by taking n = 3

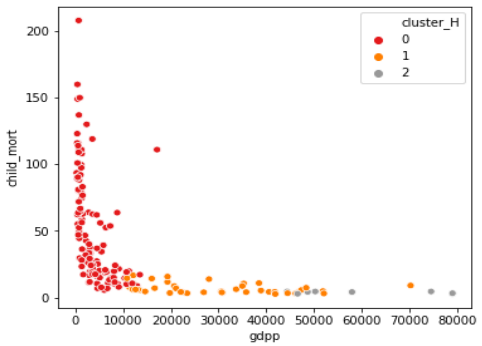












After performing the clustering using Hierarchical clustering, the clusters were formed and for visualising, I chose scatterplot and boxplot for variables : Income, GDPP and Child Mortality.

Since the size of the cluster varies significantly, I can't categorize the countries based on the level of development. However, our main focus in on cluster_H = 0 as this shows high Child Mortality and low Income and GDPP with significant number of countries in the cluster.

Clustering Profiling and Country Identification

index	cluster_H
0	118
1	41
2	8

Table 1

cluster_H	child_mort	income	gdpp
0	51.51	7581.89	3732.32
1	6.88	35853.02	29492.68
2	3.83	56321.75	60097.00

Table 2

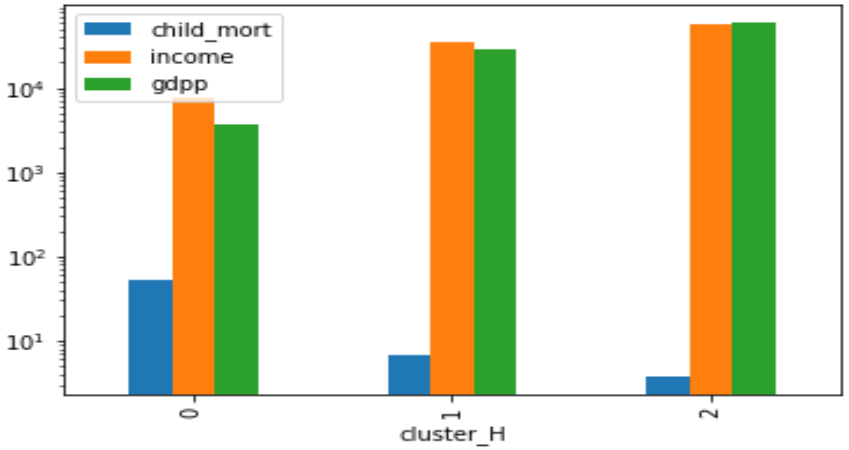


Table 1 represents the number of countries in each cluster and Table 2 represents the mean value of our main variable in each cluster.
Based on above plot and table 2, I know that countries with cluster index 0 are Under-Developed and requires aid.

Based on Hierarchical clustering, The top 5 countries that need aid are listed below :

country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	cluster_K	cluster_H
Congo, Dem. Rep.	116.00	1372740.00	264194.00	1656640.00	609.00	20.80	57.50	6.54	334.00	2	0
Liberia	89.30	624570.00	385860.00	3028020.00	700.00	5.47	60.80	5.02	327.00	2	0
Burundi	93.60	206052.00	267960.00	905520.00	764.00	12.30	57.70	6.26	231.00	2	0
Niger	123.00	772560.00	179568.00	1708680.00	814.00	2.55	58.80	6.56	348.00	2	0
Central African Republic	149.00	526280.00	177508.00	1181900.00	888.00	2.01	47.50	5.21	446.00	2	0

Conclusion and Recommendations

Although the top countries in need of aid are same by both the methods, I chose K-Means Clustering Algorithm over Hierarchical Clustering Algorithm.

- The cluster K value counts were properly divided and visualizing each cluster was possible.
- In both the methods, 3 clusters were formed but K-means gave significant plots.
- After grouping all the countries into 3 groups by using some socio-economic and health factors, I can determine the overall development of the countries.
- Here, the countries are categorised into list of developed countries, developing countries and under-developed countries.
- In Developed countries, I can see the GDP per capita and income is high where as Death of children under 5 years of age per 1000 live births i.e. child-mort is very low, which is expected.
- In Developing countries and Under-developed countries, the GDP per capita and income are low and child-mort is high.
- Specially, for under-developed countries, the death rate of children is very high, the GDP and Income are significantly low, which makes this cluster in need of driest aid.

Recommendations:

The top countries that are in need of aid are presented below:

```
Congo, Dem. Rep.
Liberia
Burundi
Niger
Central African Republic
```

The primary focus for the funding should be on countries of this segment.

- The major approach should be to provide better health facilities as this would increase the life expectancy and decrease the child mortality.
- Other focus should be on increasing the exports as that would give an increase in income and thus, GDPP will increase.
- The difference betlen the average of socio-economic factors of developed countries as compared to under-developed countries is quite huge. This funding would definitely help in improved conditions for above mentioned countries even though it might not show a significant impact