Clustering Assignment

VAIBHAV PARAKH

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

Steps Undertaken

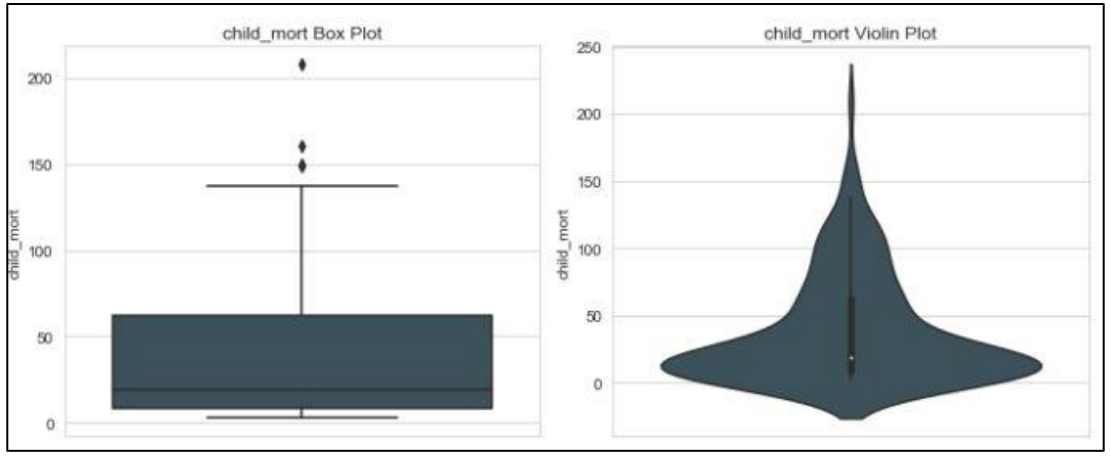
Reading and Understanding the Data	
Data Preparation	
Data Visualization	Univariate Analysis
	Bivariate Analysis
Outlier Treatment	
Re-scaling	
Modeling	KMeans Clustering
	Hierarchical Clustering
Conclusion	

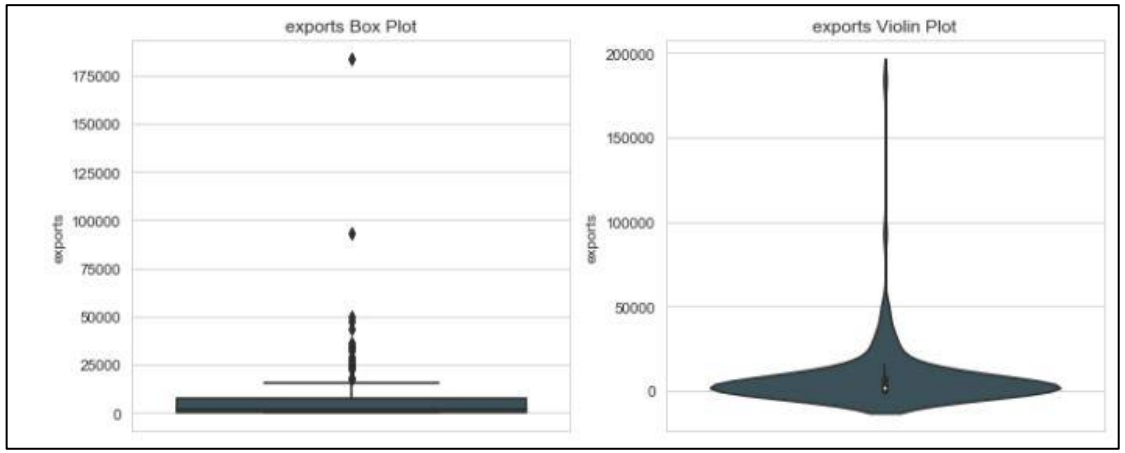
Exploratory Data Analysis

Univariate Analysis

Univariate Analysis was conducted in the form of boxpots/violinplots and distplots.

Boxplots and Violinplots where helpful in ascertaining the overall spread of the data as well as the overall density of the data.

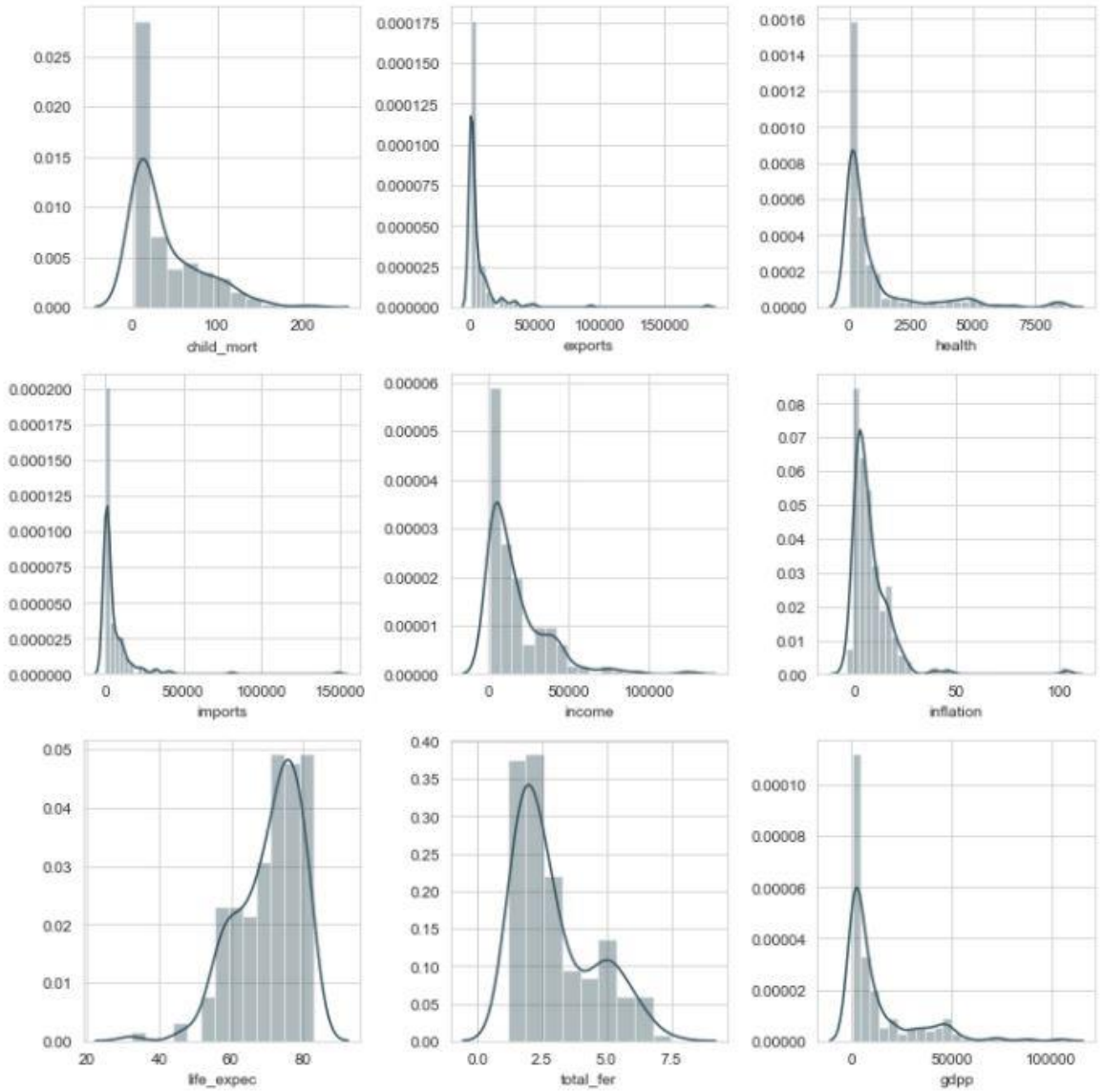




Univariate Analysis

Univariate Analysis was conducted in the form of boxpots/violinplots and distplots.

Distplots were particularly useful in looking at distribution of the datapoints of each numerical feature.

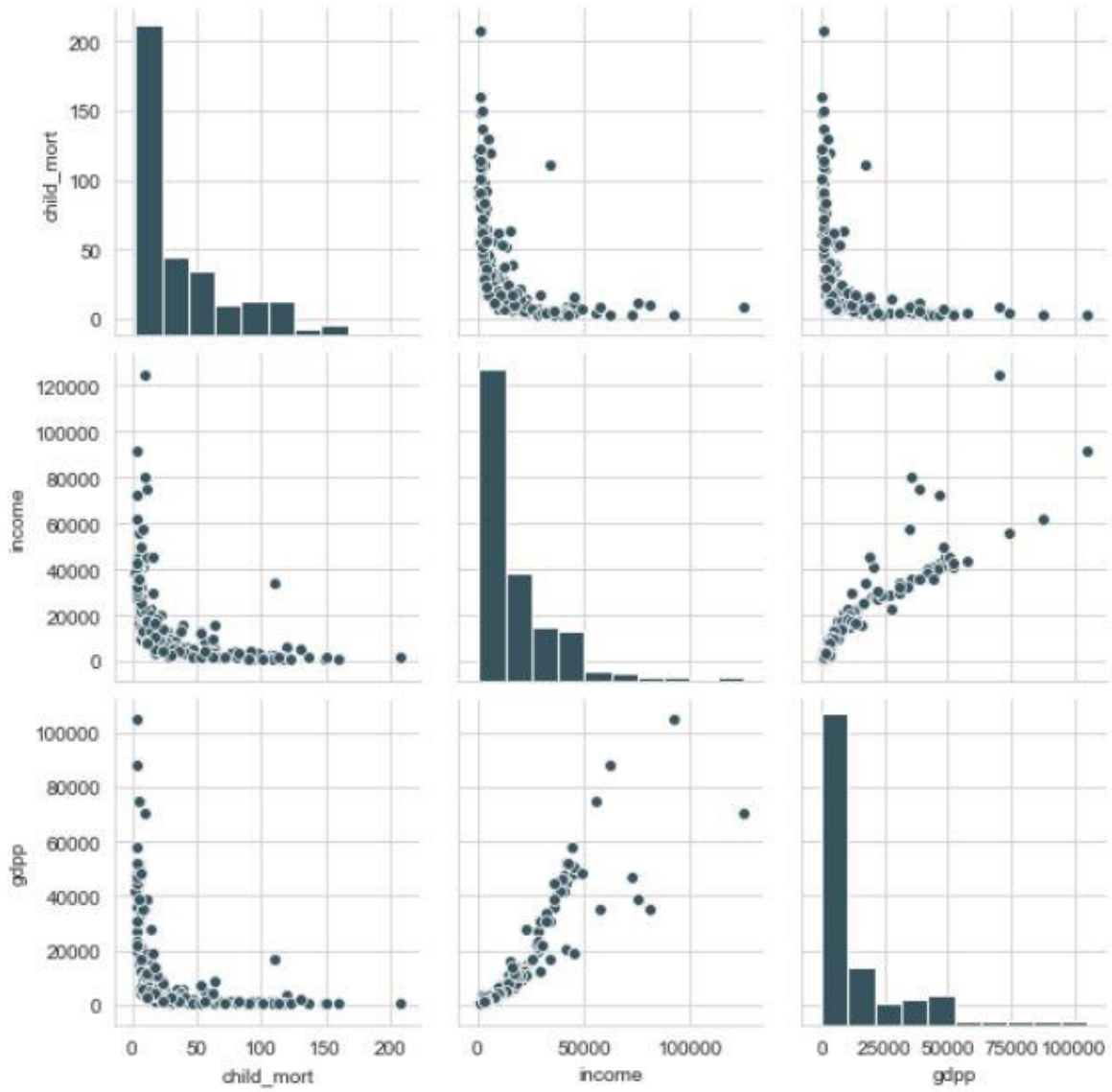


Bivariate Analysis

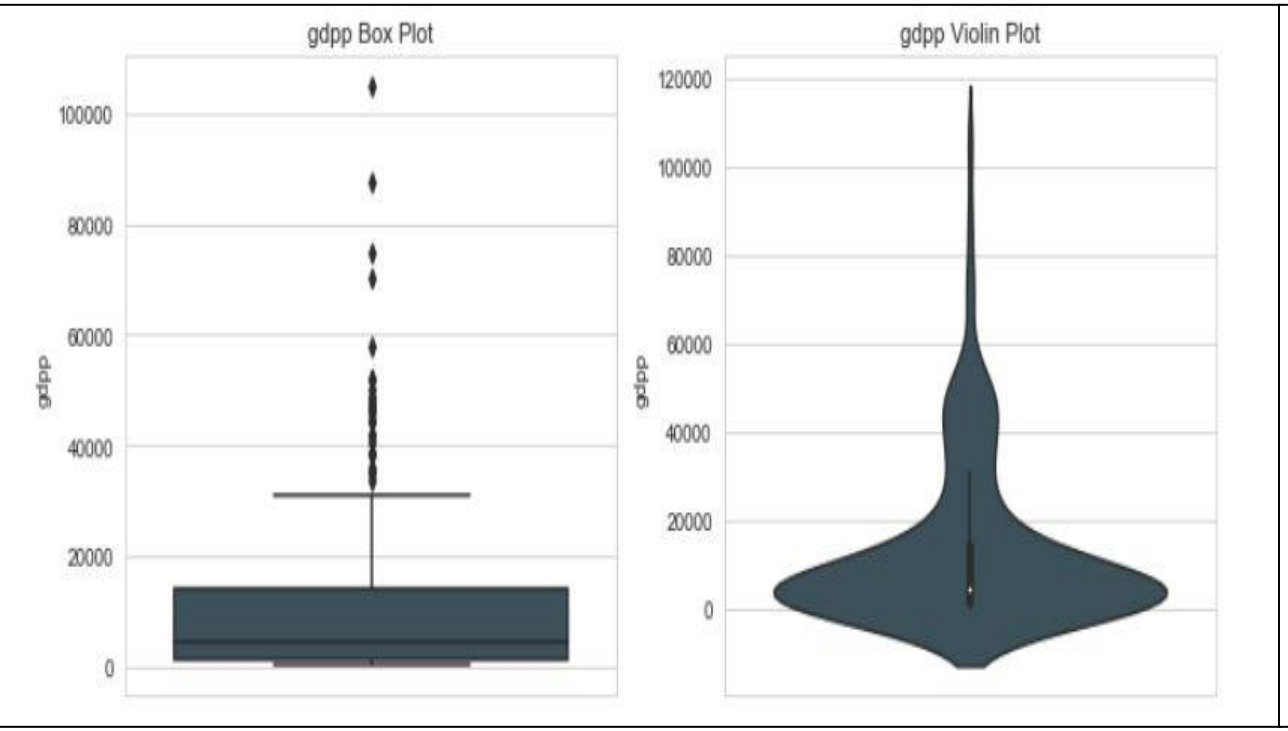
As per the business, we were advised to look at three variables:

- GDPP
- Income
- Child Mortality

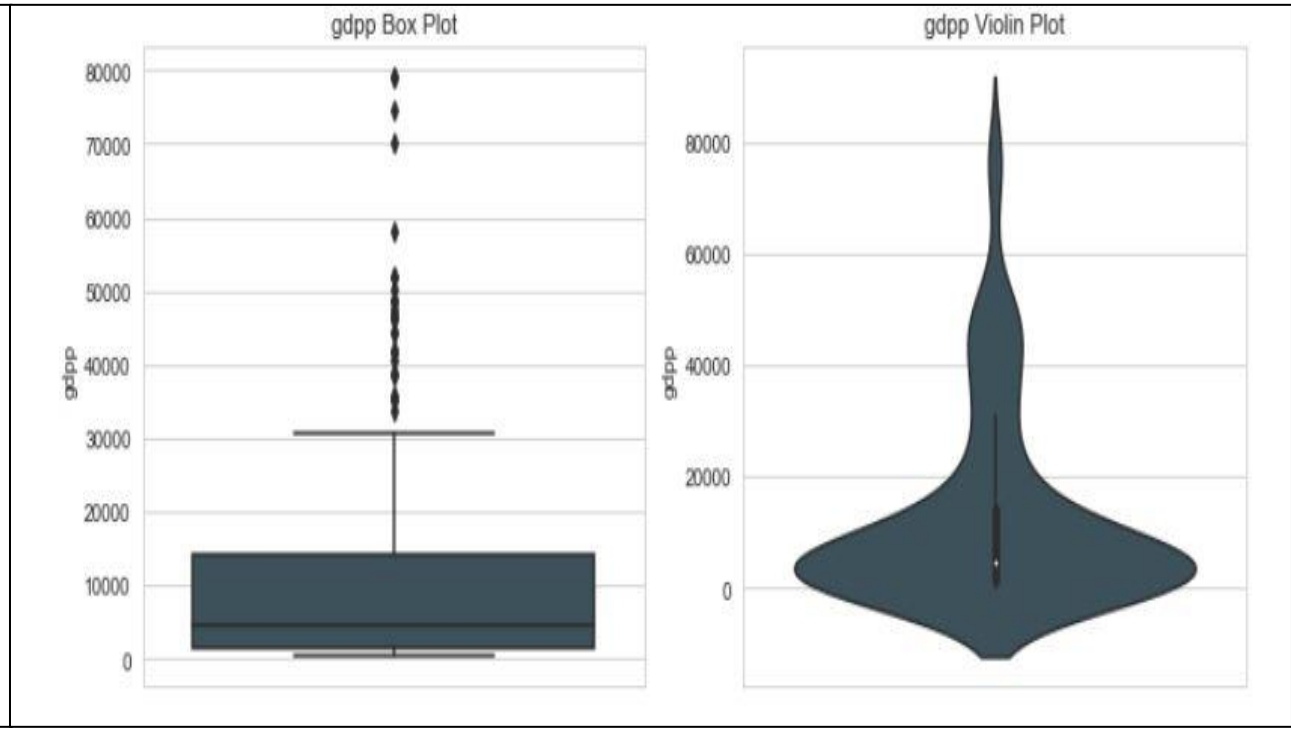
In terms of child mortality's relationship with income and GDPP, it shows a rectangular hyperbola. When income and GDPP is on the higher end of the spectrum i.e. > $ 20,000, child mortality is tending towards 0. Where as, just as income and GDPP levels dip below $20,000, child mortality sharply increases.



Outlier Treatment

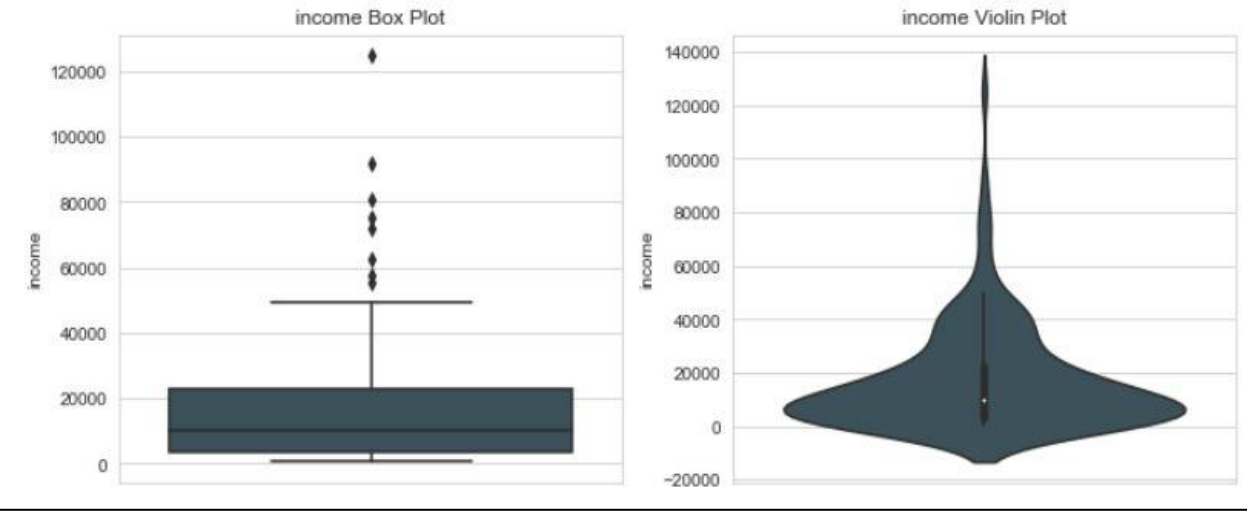


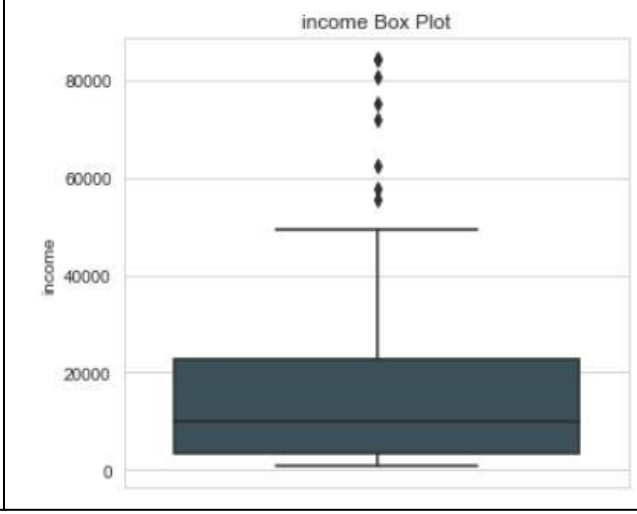
Pre-Outlier Treatment

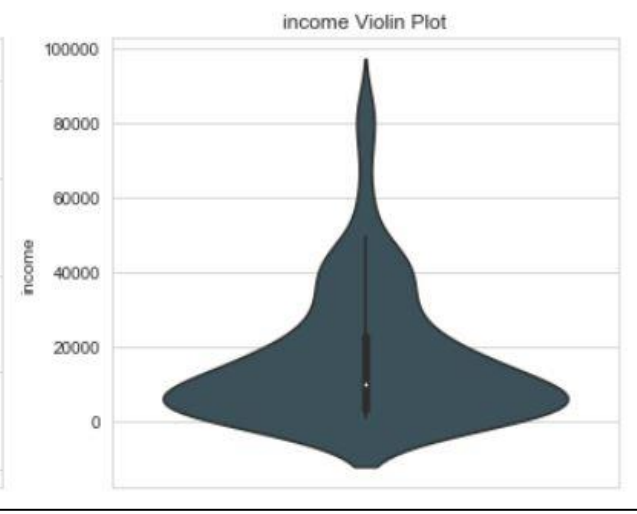


Post-Outlier Treatment

For outlier treatment, we floored and capped values at 1%ile and 99%ile, respectively. This ensured that we have a smoother distribution and at the same time we include all 167 countries and do not leave out any country from clustering exercise. Presented are distribution of few numerical features pre- and post-outlier treatment.

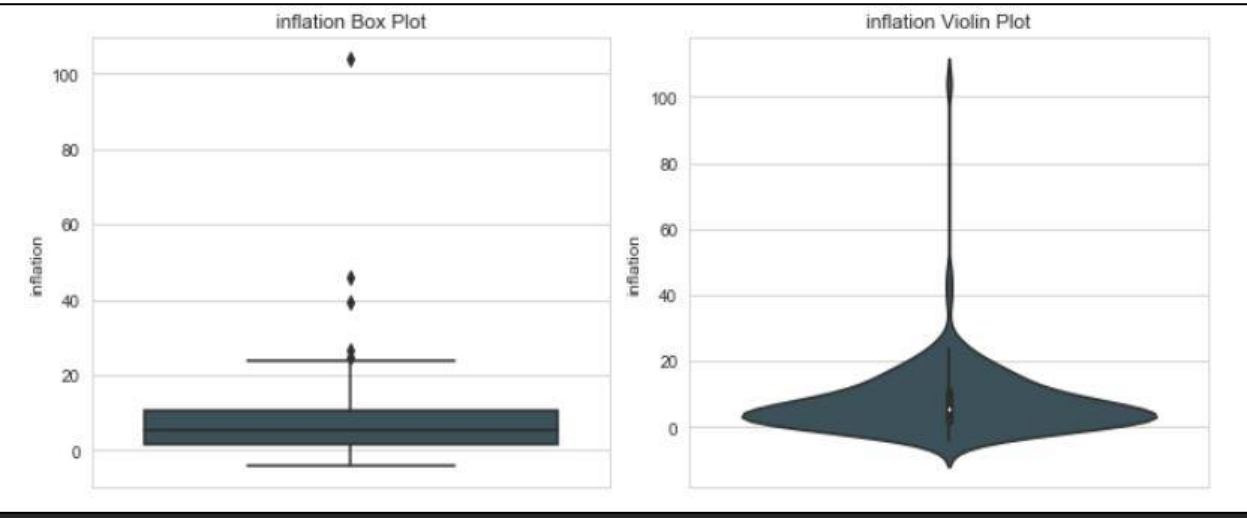


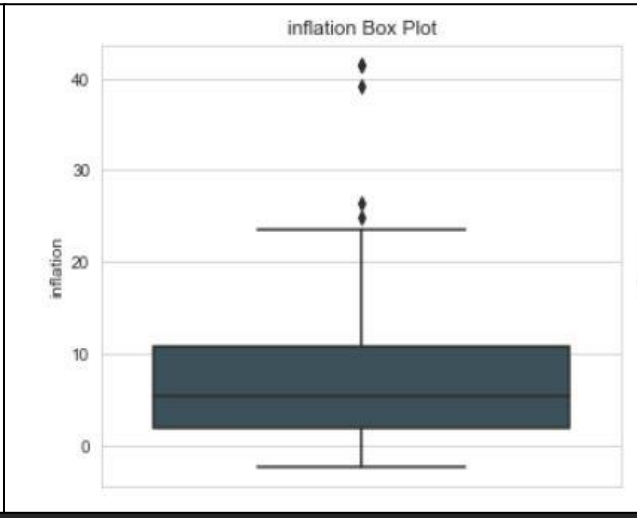


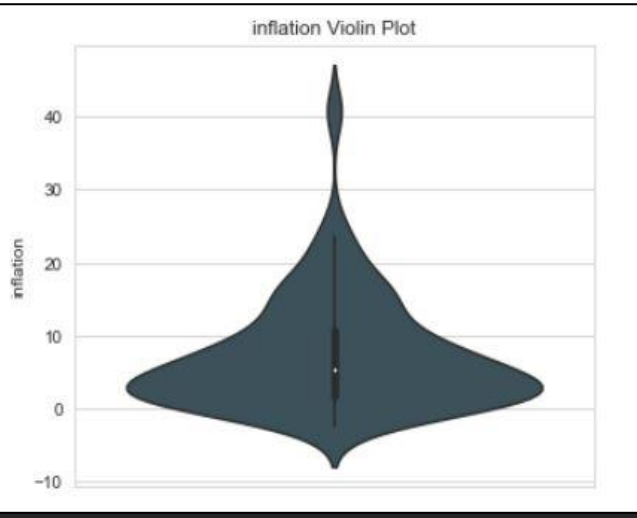


Pre-Outlier Treatment

Post-Outlier Treatment





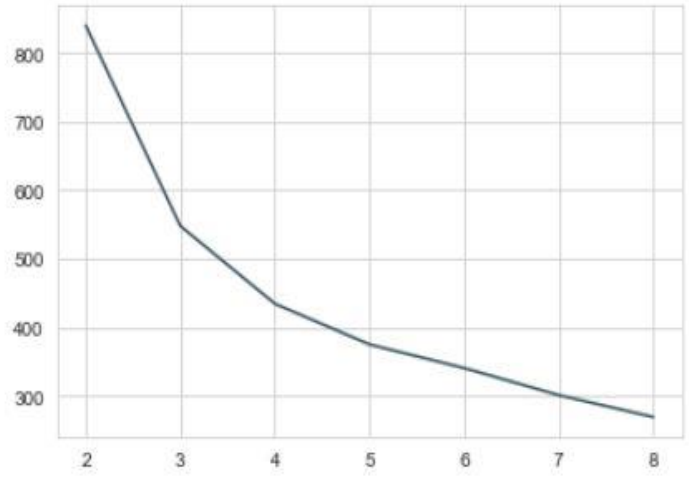


Hopkins Statistic

Value came out to be 0.9119 which indicates that the data is highly clustered .

K-Means Clustering

```
plt.plot('Number of Clusters', 'SSD Value', data = ssd)
plt.show()
```

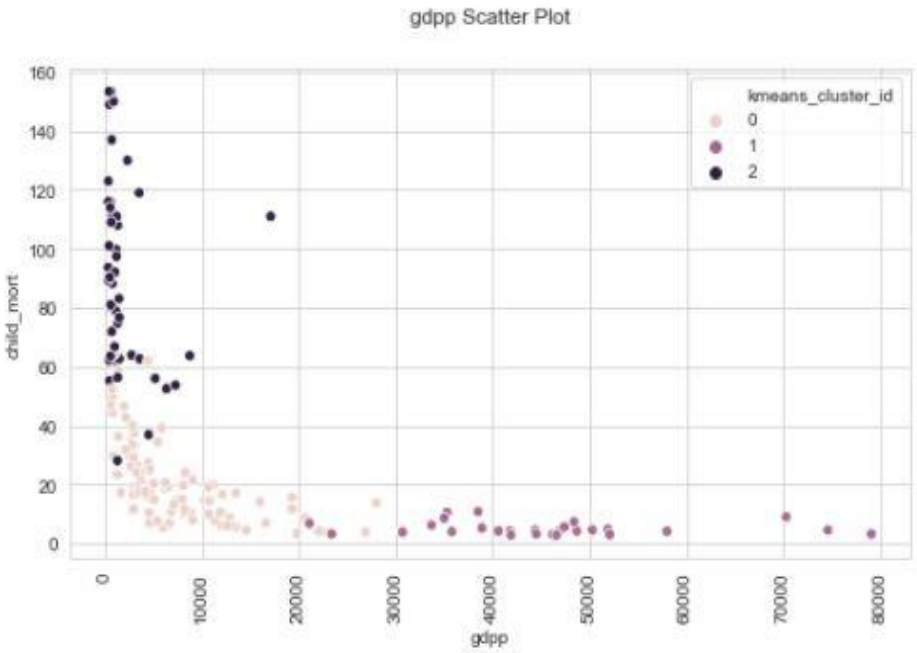


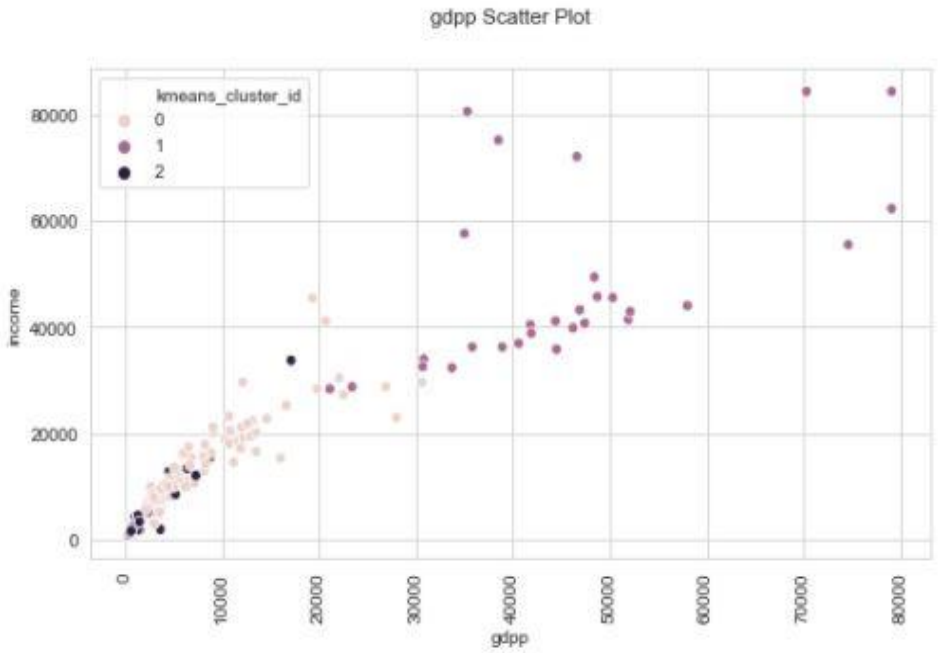
While implementing K-Means, we looked at sum of squared distances values and from the graph we could see SSD value at number of clusters = 3 is the sharpest corner of the graph.

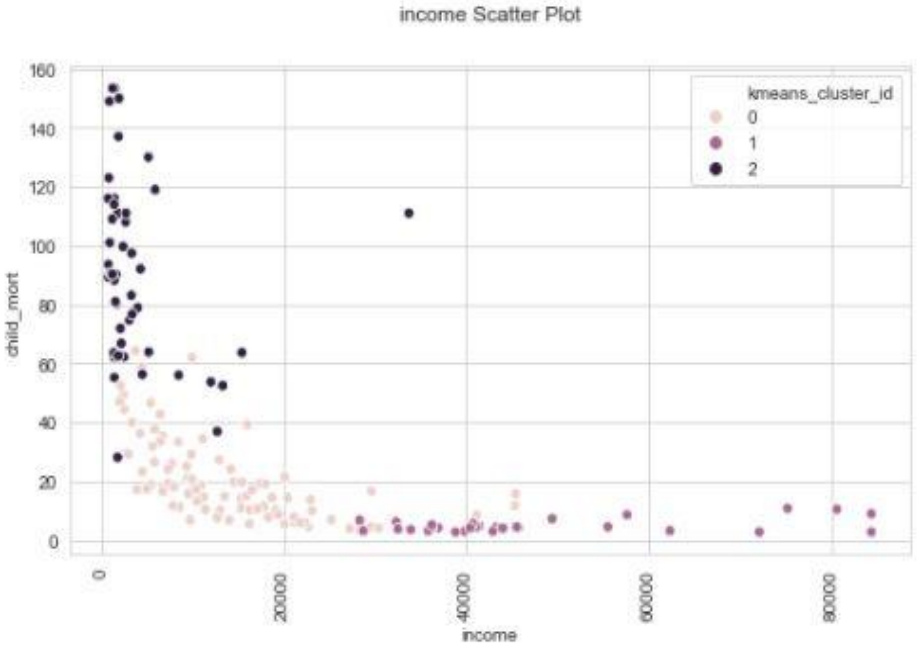
Also, we calculated the average silhouette scores and from that we again narrowed in on 3 clusters as preferred number of clusters for classifying all 167 countries into.

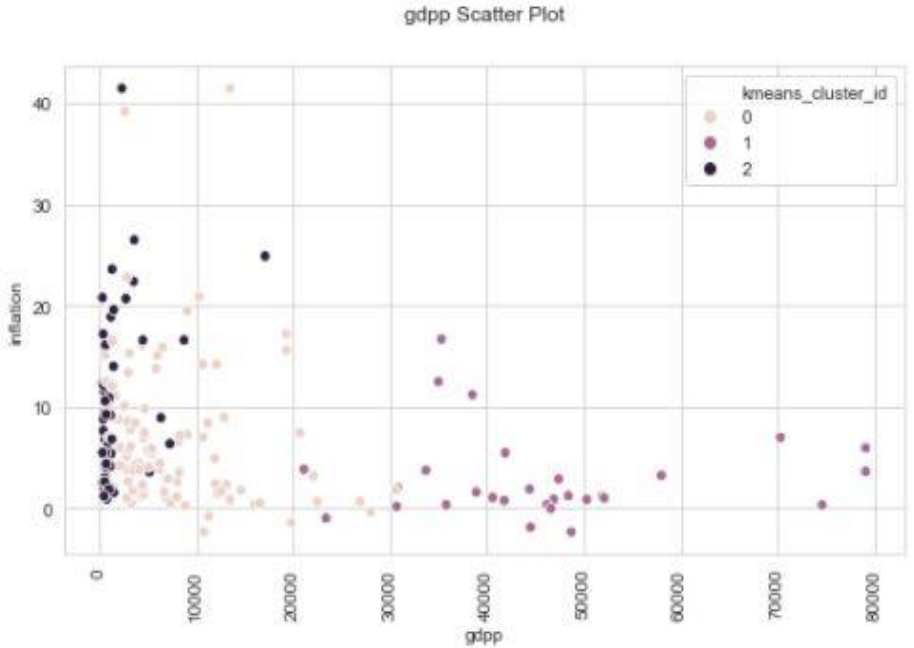
```
# silhouette analysis
range_n_clusters = [2, 3, 4, 5, 6, 7, 8]

for num_clusters in range_n_clusters:

    # intialise kmeans
    kmeans = KMeans(n_clusters=num_clusters, max_iter=50)
    kmeans.fit(df_1)

    cluster_labels = kmeans.labels_

    # silhouette score
    silhouette_avg = silhouette_score(df_1, cluster_labels)
    print("For n_clusters={0}, the silhouette score is {1}".format(num_clusters, silhouette_avg))
```

```
For n_clusters=2, the silhouette score is 0.46939980287788113
For n_clusters=3, the silhouette score is 0.40708993455880516
For n_clusters=4, the silhouette score is 0.39539142309551445
For n_clusters=5, the silhouette score is 0.3864288935632213
For n_clusters=6, the silhouette score is 0.2757417377396905
For n_clusters=7, the silhouette score is 0.2932516338403447
For n_clusters=8, the silhouette score is 0.3109403883958476
```





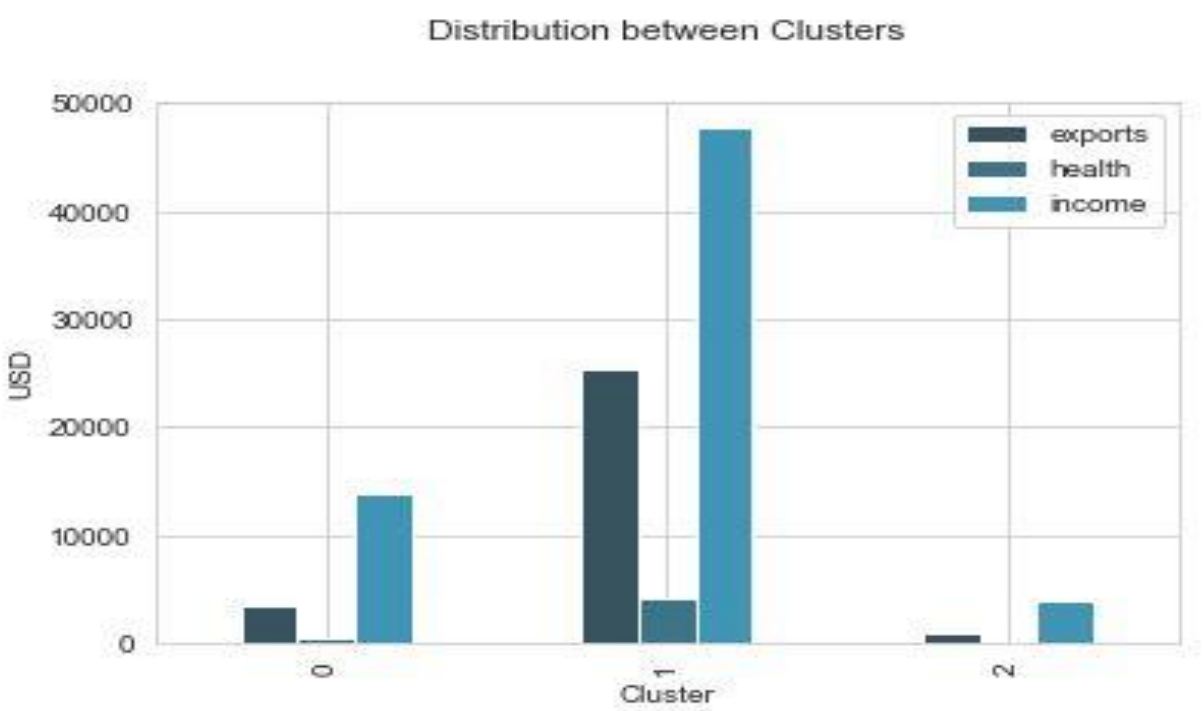




From the following 4 plots, we could see that cluster 2 is coming out to be the cluster of interest.

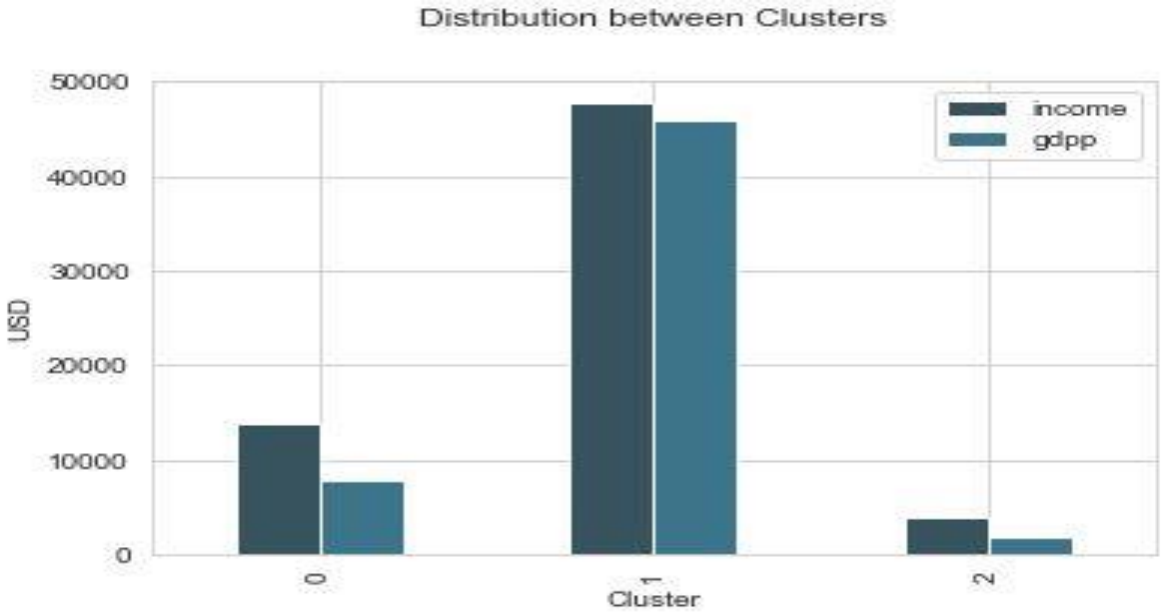
(Anti-clock-wise) cluster labelled 2 countries represent very high child mortality rate (from 60 to up to 160) and very low GDPP figures (tending to 0) and income figures in the first plot and third plot.

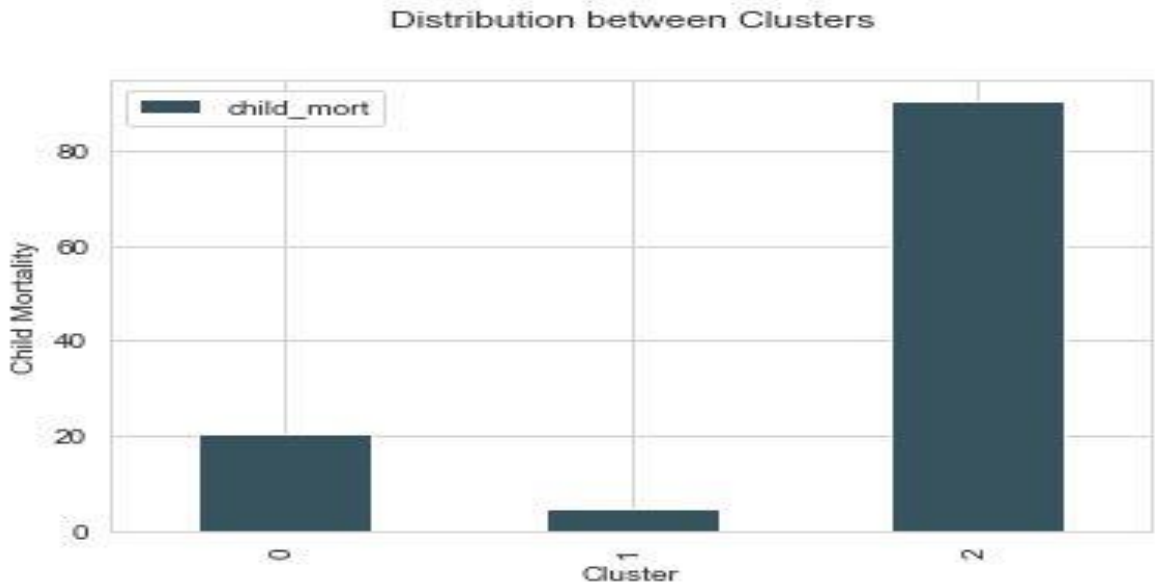
In the income vs GDPP plot, cluster 2 countries represent the bottom countries which also seem to have a very high inflation as evident from the fourth plot .



These graphs were also very helpful in terms of helping the analyst pitch which cluster of countries to focus on. Clearly, cluster 2 labelled countries need the most help which is evident from very low exports, expenditure on health, income and GDPP figures from top and bottom left graphs.

Also, bottom right informs us that cluster 2 labelled countries have the highest child mortality rate.





Top 5 Countries as per K-Means

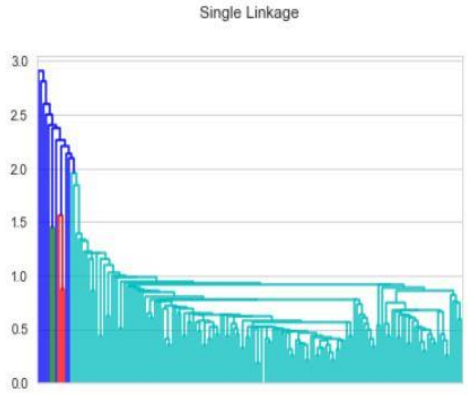
	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	Mean
39	Solomon Islands	-2.039563	-0.109081	-0.028367	0.144390	-0.383580	-0.451031	0.409237	-0.743222	-0.214717	-0.879287
14	Eritrea	-1.151449	-0.384055	-0.598176	-0.468958	-0.448686	0.117714	0.409237	-0.367670	-0.493912	-0.698015
27	Madagascar	-0.922047	-0.348089	-0.598176	-0.426141	-0.454111	-0.215934	0.236556	-0.377820	-0.517754	-0.631304
36	Rwanda	-0.876166	-0.364102	-0.340985	-0.431840	-0.461345	-0.949722	0.965652	-0.469170	-0.465923	-0.601145
22	Kenya	-0.922047	-0.304606	-0.421507	-0.329517	-0.256987	-1.011465	0.620291	-0.611271	-0.326326	-0.501786

As per K-Means, the top 5 countries which are in dire need of immediate aid are:

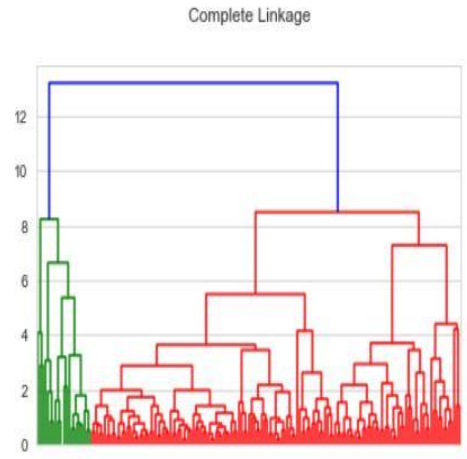
1. Solomon Islands
2. Eritrea
3. Madagascar
4. Rwanda
5. Kenya

Hierarchical Clustering

```python
# Single Linkage

single_linkage = linkage(df_2, method="single", metric='euclidean')
dendrogram(single_linkage)
plt.title('Single Linkage')
plt.xticks(ticks=[])
plt.show()
```

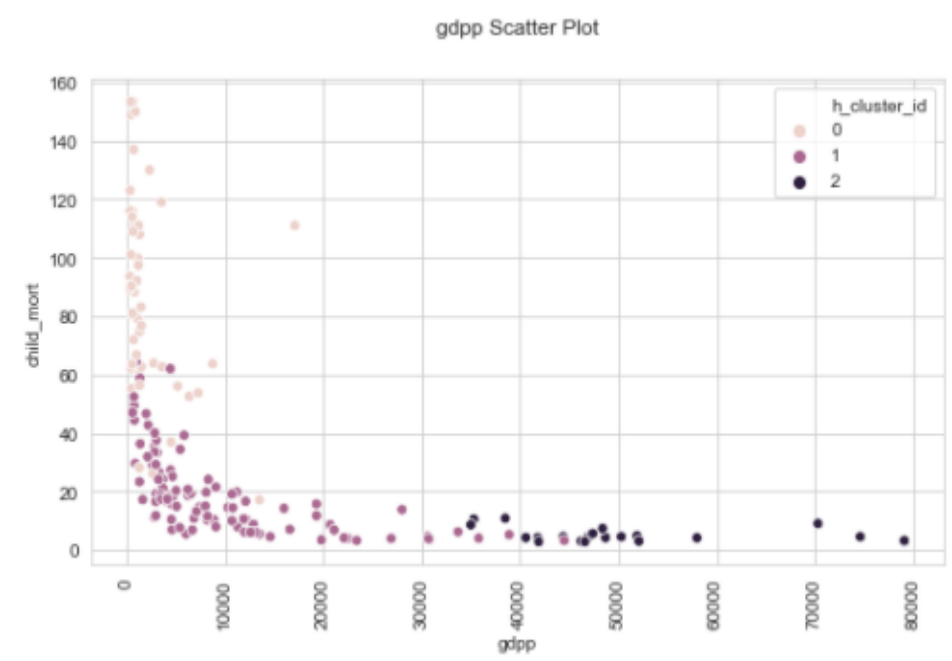


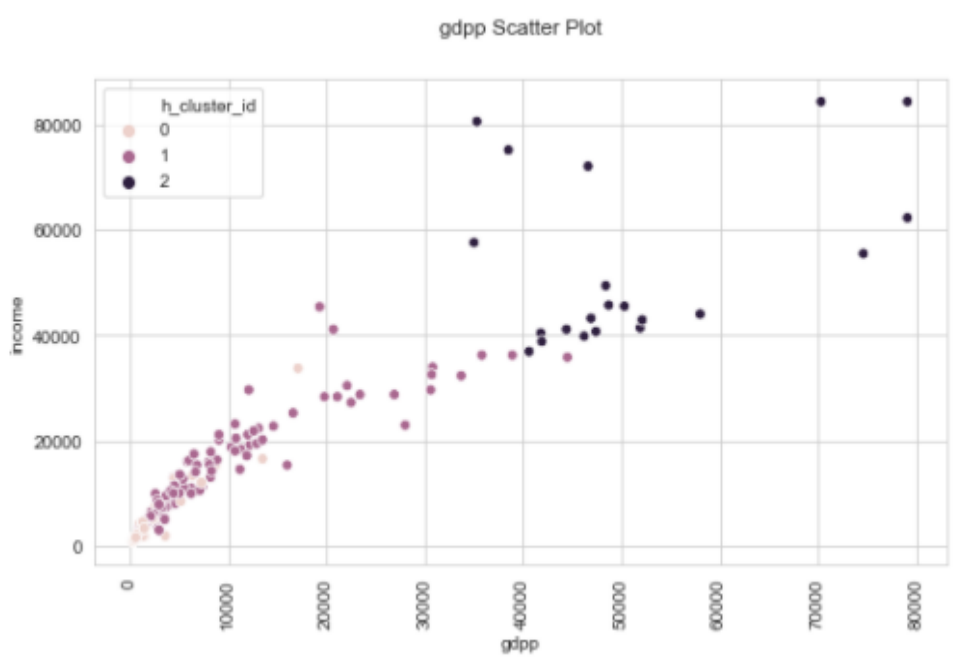

```python
# Complete Linkage

complete_linkage = linkage(df_2, method="complete", metric='euclidean')
dendrogram(complete_linkage)
plt.title('Complete Linkage')
plt.xticks(ticks=[])
plt.show()
```

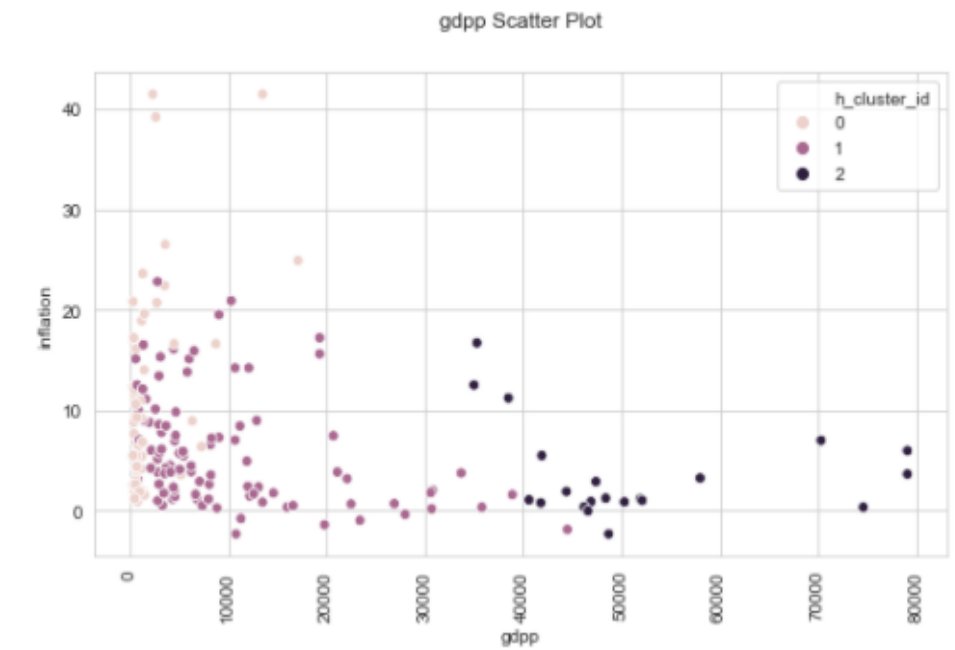


While performing Hierarchical Clustering, single and complete linkage methods were made use of to form the dendrogram. Single linkage dendrogram did not make sense.

Complete Linkage dendrogram on the contrary gave us an easier to interpret graph. I chose to cut the tree at 3 clusters as also informed by choosing 3 clusters as my ideal number if clusters from K-Means to compare if we do get the same results or different results.







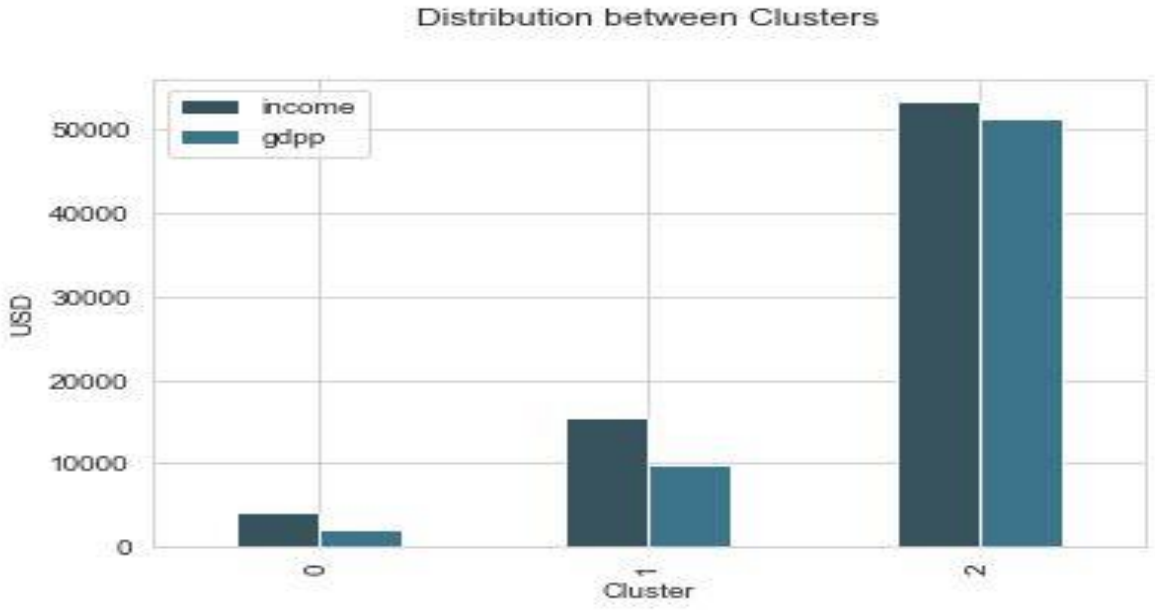
From the following 4 plots, we could see that cluster 0 is coming out to be the cluster of interest.

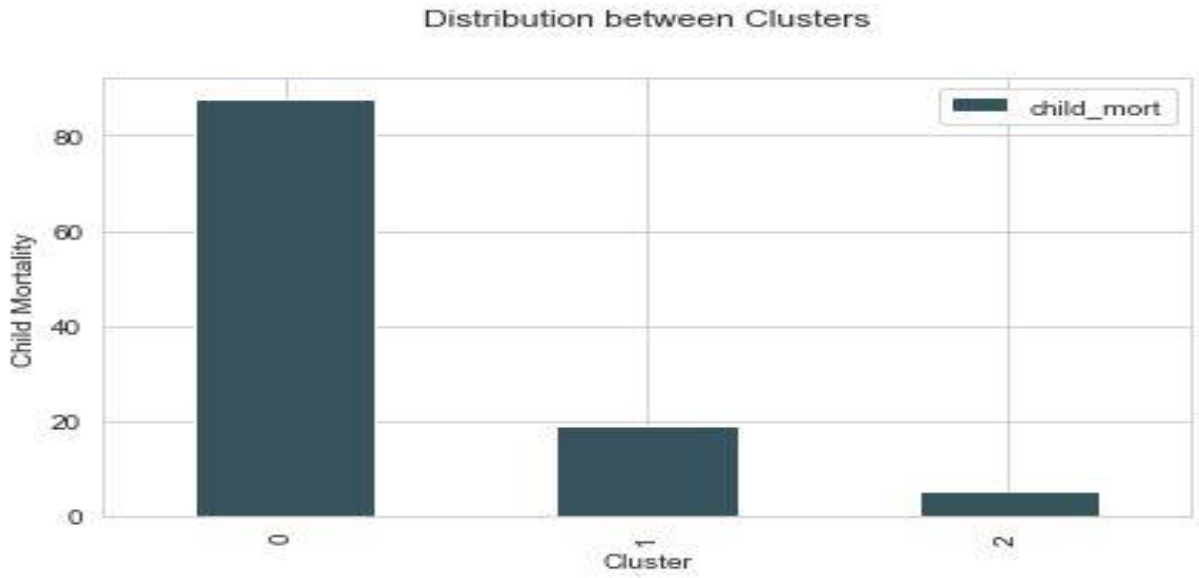
(Anti-clock-wise) cluster labelled 0 countries represent very high child mortality rate (from 60 to up to 160) and very low GDPP figures (tending to 0) and income figures in the first plot and third plot.

In the income vs GDPP plot, cluster 0 countries represent the bottom countries which also seem to have a very high inflation as evident from the fourth plot .

Clearly, cluster 0 labelled countries need the most help which is evident from very low exports, expenditure on health, income and GDPP figures from top and bottom left graphs.

Also, bottom right informs us that cluster 0 labelled countries have the highest child mortality rate.





Top 5 Countries as per Hierarchical Clustering

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	Mean
40	Solomon Islands	-1.813380	-0.139275	-0.091031	0.116169	-0.428234	-0.493657	0.298700	-0.590984	-0.265630	-0.835748
14	Eritrea	-0.987259	-0.414925	-0.615653	-0.502239	-0.491180	-0.019579	0.298700	-0.246926	-0.512915	-0.663785
27	Madagascar	-0.773870	-0.378871	-0.615653	-0.459068	-0.496425	-0.297691	0.139024	-0.256225	-0.534032	-0.601443
37	Rwanda	-0.731193	-0.394923	-0.378859	-0.464814	-0.503419	-0.909341	0.813210	-0.339915	-0.488125	-0.574246
22	Kenya	-0.773870	-0.335281	-0.452995	-0.361647	-0.305840	-0.960807	0.493859	-0.470099	-0.364483	-0.481398

As per Hierarchical clustering, the top 5 countries which are in dire need of immediate aid are:

1. Solomon Islands
2. Eritrea
3. Madagascar
4. Rwanda
5. Kenya

Conclusion

As per both K-Means and Hierarchical clustering, the top 5 countries which are in dire need of immediate aid are:

1. Solomon Islands
2. Eritrea
3. Madagascar
4. Rwanda
5. Kenya