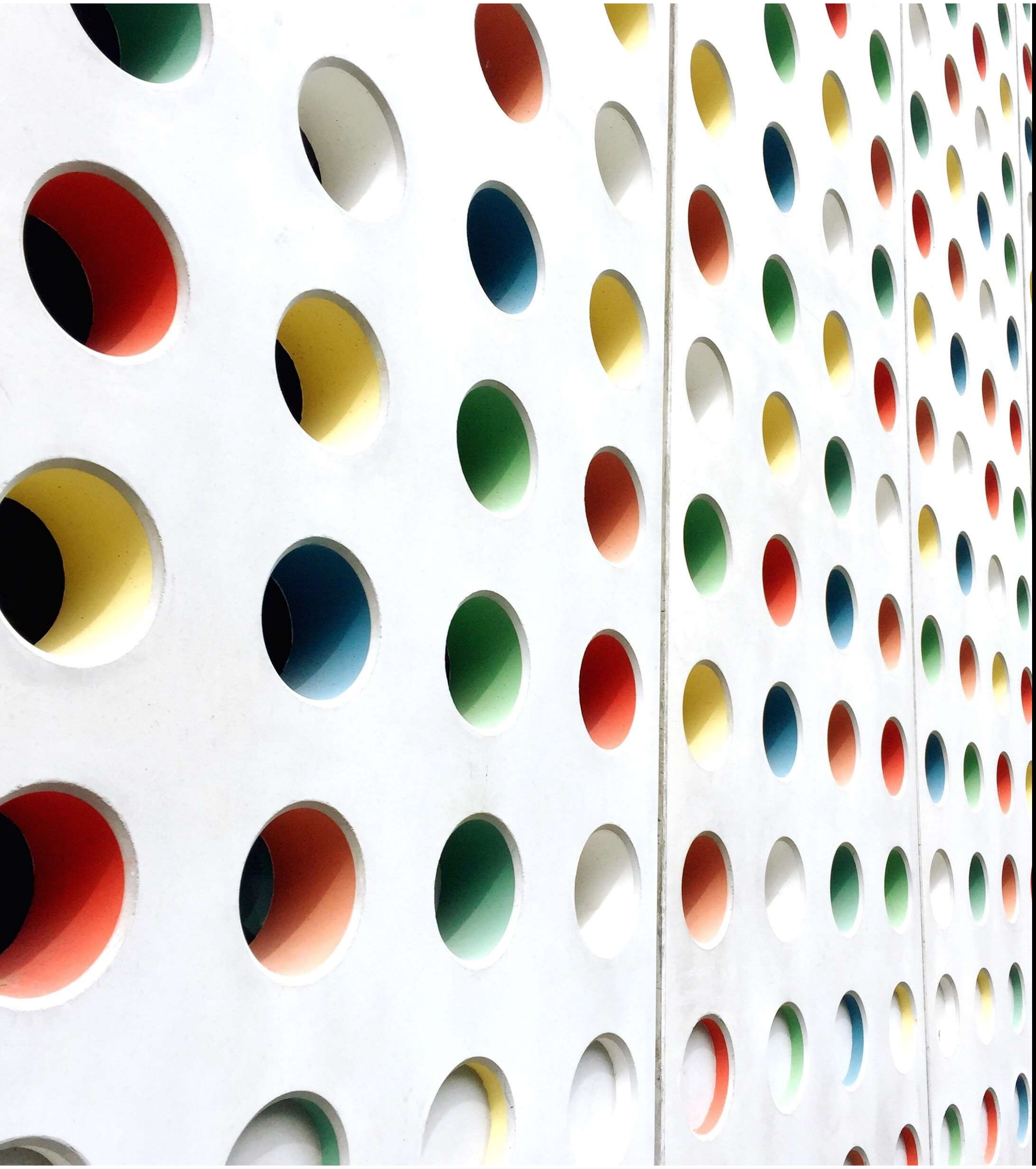
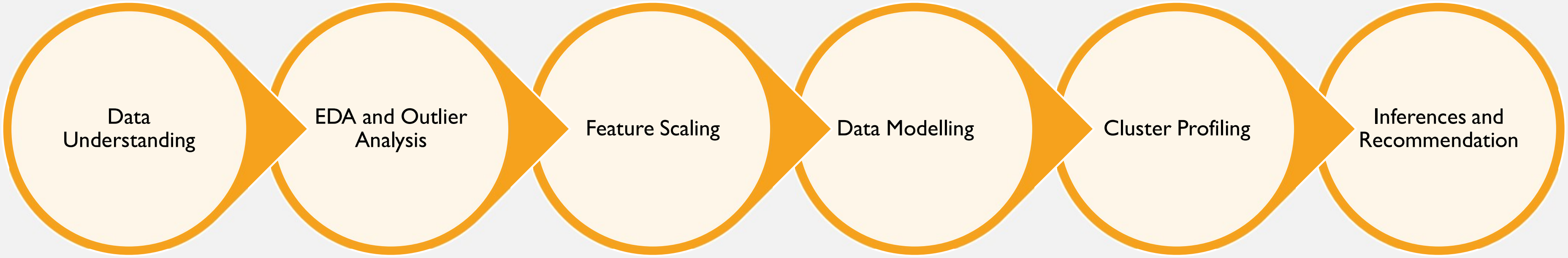


CLUSTERING ASSIGNMENT

- Harsha Belurkar

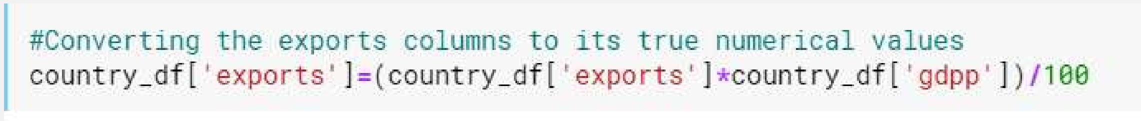
PROBLEM STATEMENT AND ITS APPROACH

- 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 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.
- <u>Model Objective:</u>
- Cluster the countries by the factors mentioned and recommend to the CEO.

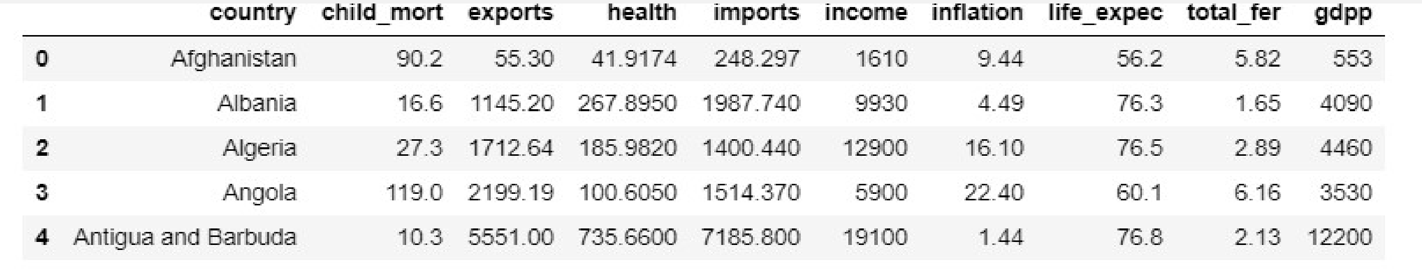


DATA UNDERSTANDING

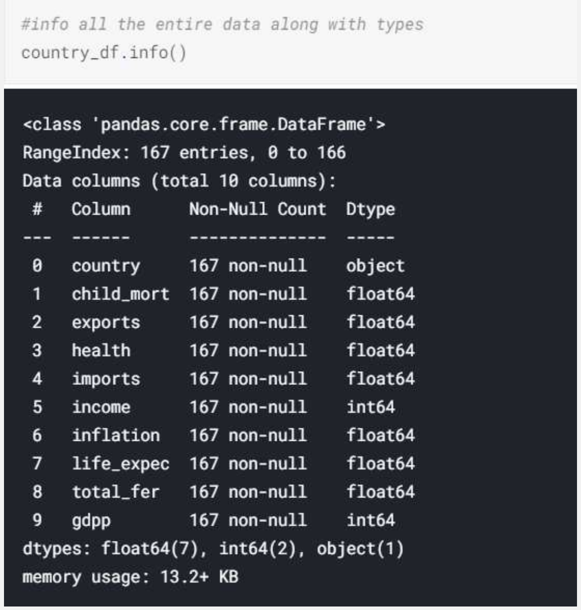
- Data set used: country_data.csv
- Have imported necessary libraries required.
- Inspected data using Shape, Info, Describe and head functions
- Shape of the data frame: (167,10)
- Total number of columns with missing values : 0
- After verifying the data dictionary our observation :The features exports, health and imports are given as percentage of GPA. Therefore I have converted it into its true numerical values.



```
#Converting the exports columns to its true numerical values
country_df['exports']=(country_df['exports']*country_df['gdpp'])/100
```



	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

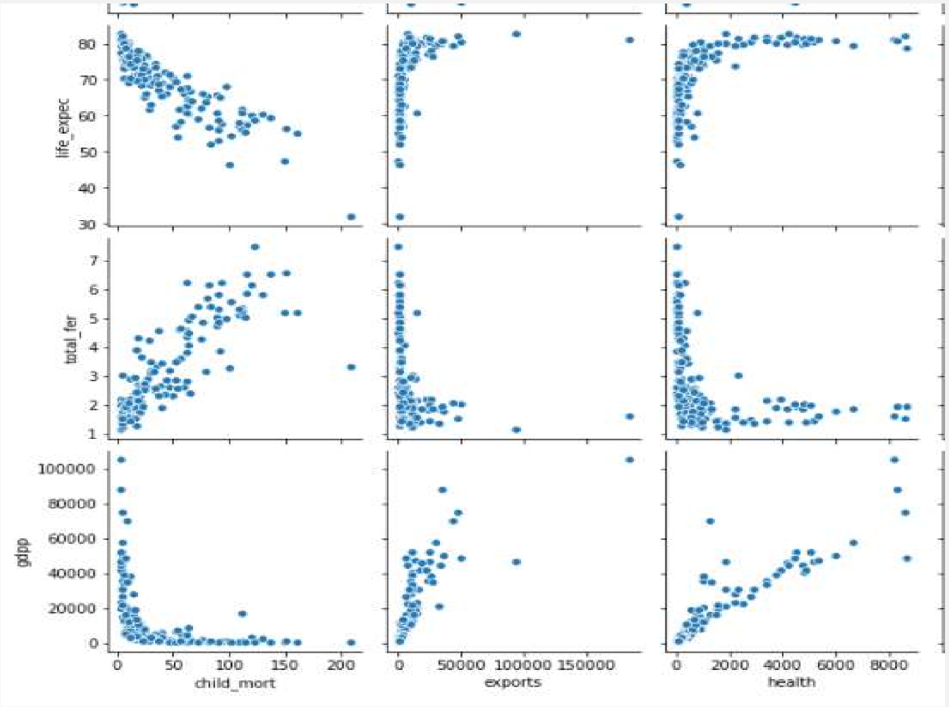


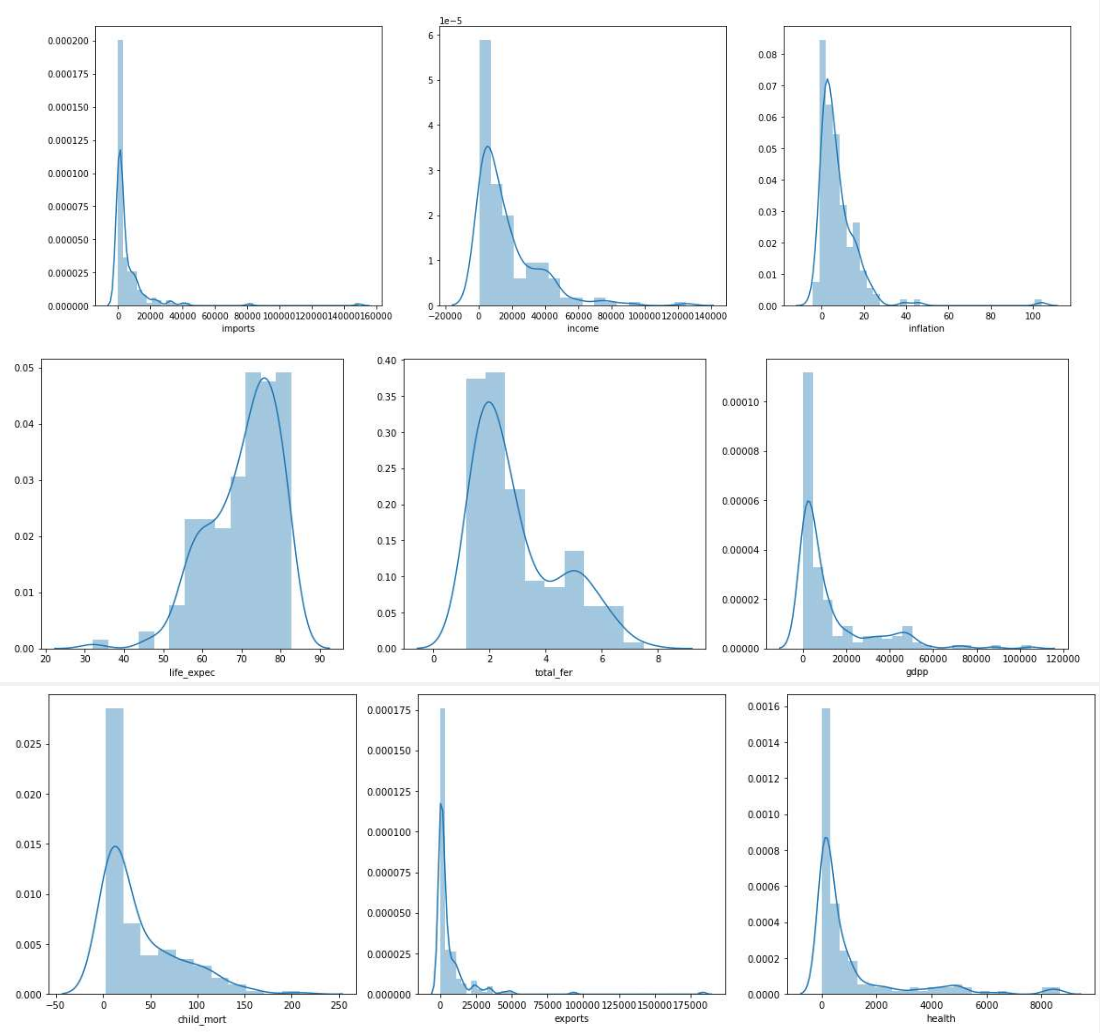
```
#info all the entire data along with types
country_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 167 entries, 0 to 166
Data columns (total 10 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   country     167 non-null    object
 1   child_mort  167 non-null    float64
 2   exports     167 non-null    float64
 3   health      167 non-null    float64
 4   imports     167 non-null    float64
 5   income      167 non-null    int64
 6   inflation   167 non-null    float64
 7   life_expec  167 non-null    float64
 8   total_fer   167 non-null    float64
 9   gdpp        167 non-null    int64
dtypes: float64(7), int64(2), object(1)
memory usage: 13.2+ KB
```

EXPLORATORY DATA ANALYSIS

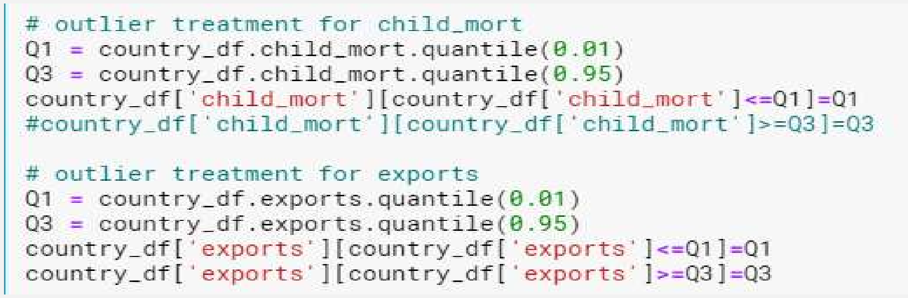
- Seaborn Distplot represents the overall distribution of continuous data variables.
- The above graphs shows the features child_mortality , gdpp , income, total_fer ,life_expec shows are widely distributed.





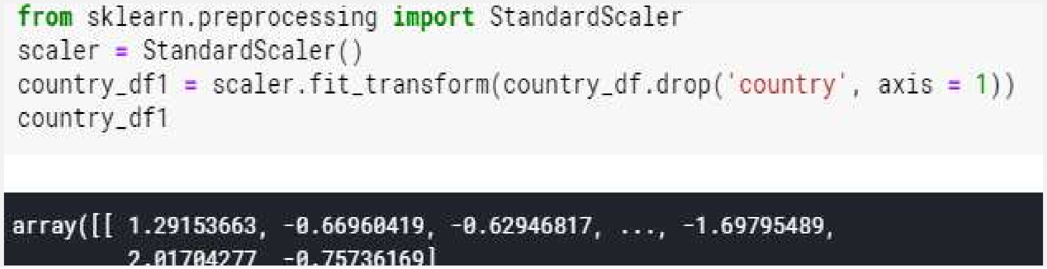
OUTLIER ANALYSIS AND SCALING

- Observation: We find that all the features have outliers. We need to handle them expect the outliers which are in higher range of child mortality and lower range in gdpp and income. As there factors help us in cluster profiling as the countries if these three characteristics may require the funding more. As the data which is provided is less we chose not to delete any outliers and instead we use capping method.



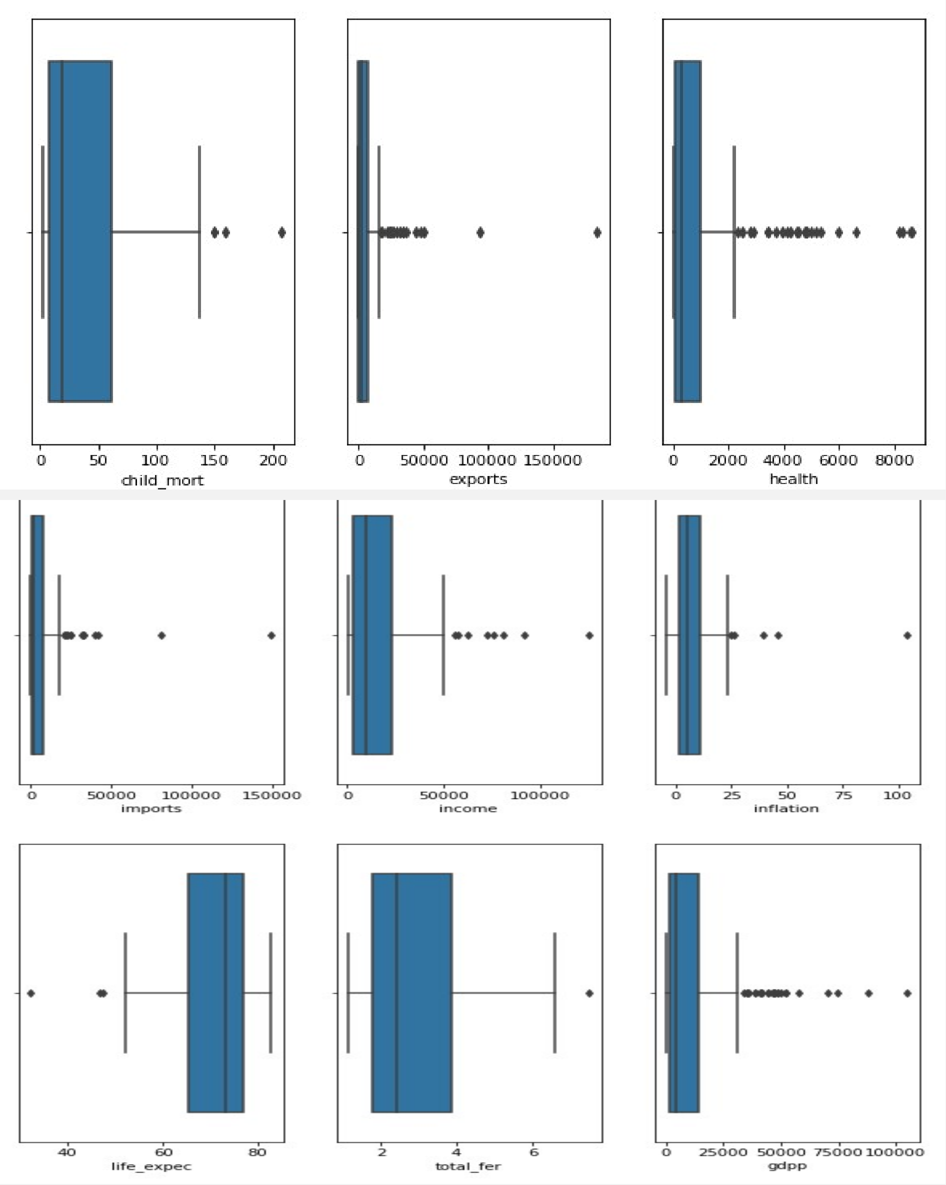
```
# outlier treatment for child_mort
Q1 = country_df.child_mort.quantile(0.01)
Q3 = country_df.child_mort.quantile(0.95)
country_df['child_mort'][country_df['child_mort']<=Q1]=Q1
#country_df['child_mort'][country_df['child_mort']>=Q3]=Q3

# outlier treatment for exports
Q1 = country_df.exports.quantile(0.01)
Q3 = country_df.exports.quantile(0.95)
country_df['exports'][country_df['exports']<=Q1]=Q1
country_df['exports'][country_df['exports']>=Q3]=Q3
```

Features are scaled using Standard scaler



```
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
country_df1 = scaler.fit_transform(country_df.drop('country', axis = 1))
country_df1
```

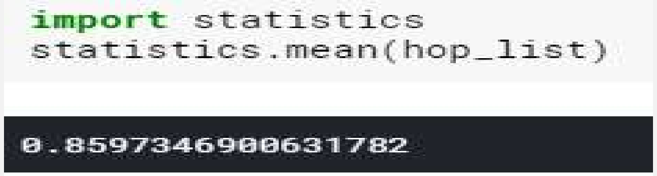
```
array([[ 1.29153663, -0.66960419, -0.62946817, ..., -1.69795489,
         2.01704277, -0.75736169]
```



DATA PROCESSING

Hopkins Stats: If the value of Hopkins statistic is close to 1, then conclude that the dataset D is significantly a cluster able data. As Hopkins stats value changes every time we run the code. So I prefer to take the avg of them and then see the score.

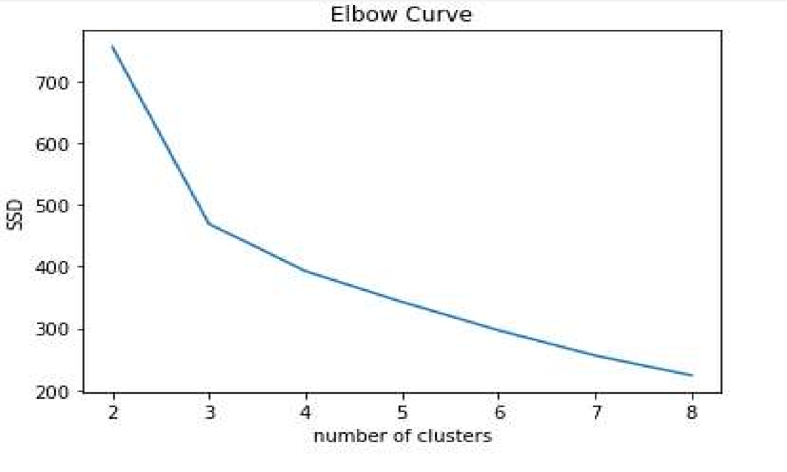
The Hopkins score is about 0.85. IT denotes that the data is good for clustering



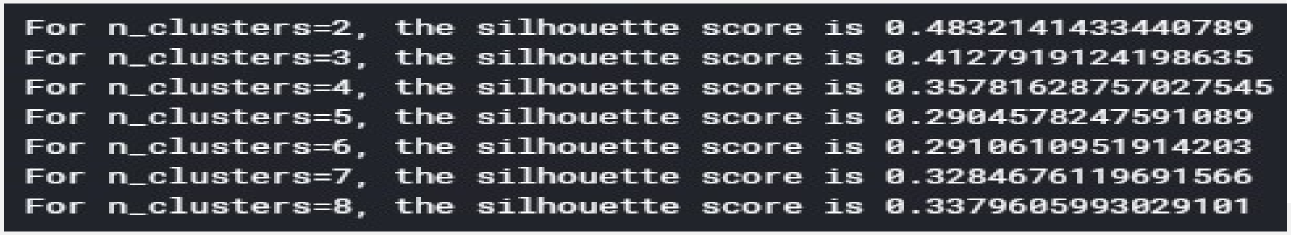
```
import statistics
statistics.mean(hop_list)
```

```
0.8597346900631782
```

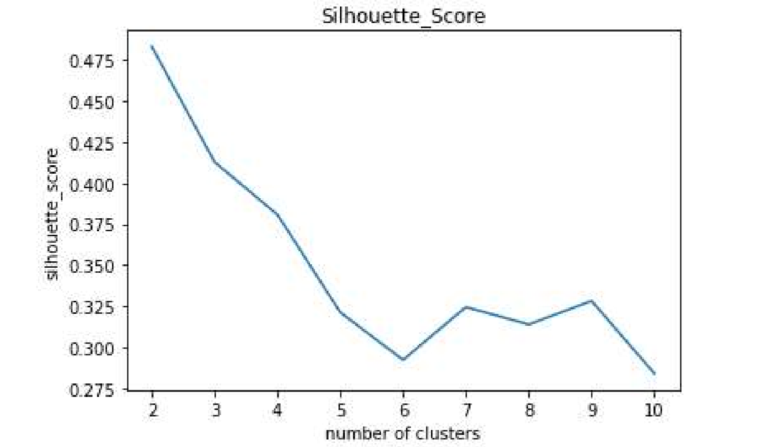
Elbow curve: The graph shows that cluster size 3 is optimal. ssd is sum of square distances to individual samples to the nearest clusters. As n of clusters increases ssd will decrease. we need to find does it goes down enough to add one cluster from 3 to 4 slope reduces. so the drop is not the significant. So optimal value is 3.



Silhouette score : The graph and the silhouette_score shows that k=2 or k=3 are optimal values and can be used as number of clusters.

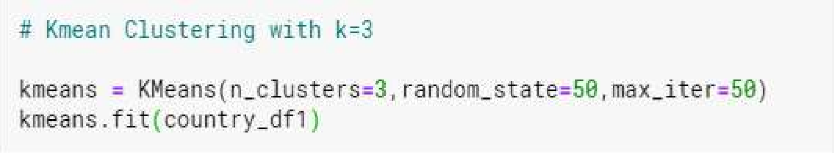


```
For n_clusters=2, the silhouette score is 0.4832141433440789
For n_clusters=3, the silhouette score is 0.4127919124198635
For n_clusters=4, the silhouette score is 0.35781628757027545
For n_clusters=5, the silhouette score is 0.2904578247591089
For n_clusters=6, the silhouette score is 0.2910610951914203
For n_clusters=7, the silhouette score is 0.3284676119691566
For n_clusters=8, the silhouette score is 0.3379605993029101
```

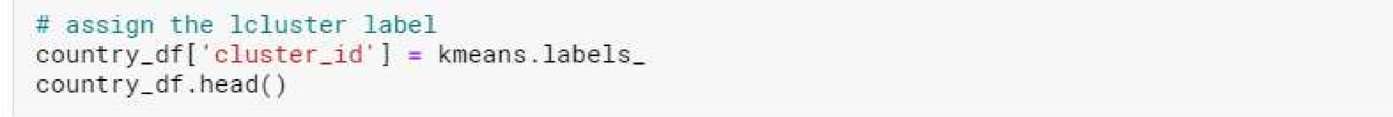


DATA MODELING – K MEANS MODEL

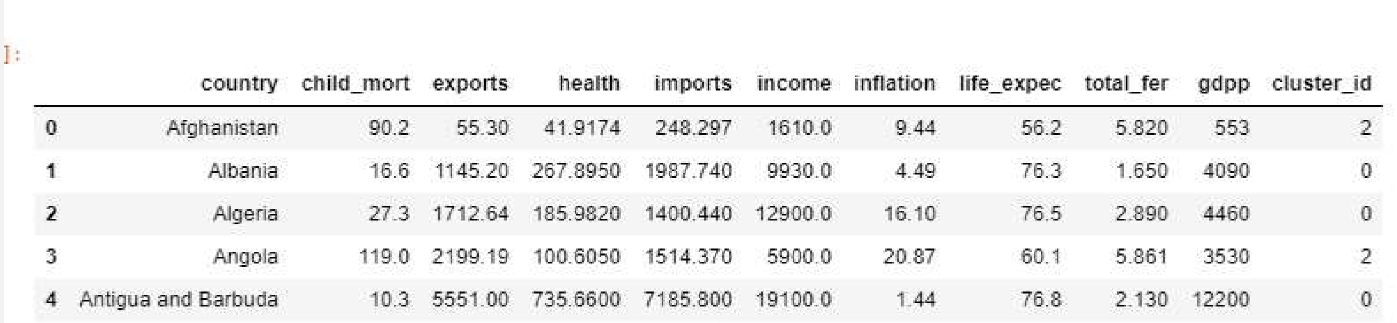
- K Value taken is 3



```
# Kmean Clustering with k=3

kmeans = KMeans(n_clusters=3,random_state=50,max_iter=50)
kmeans.fit(country_df1)
```

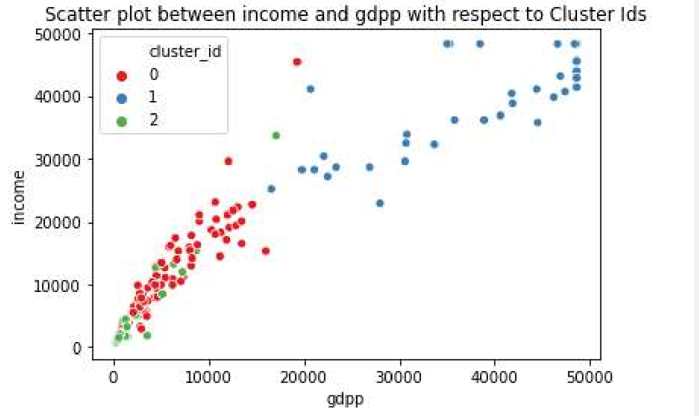


```
# assign the lcluster label
country_df['cluster_id'] = kmeans.labels_
country_df.head()
```

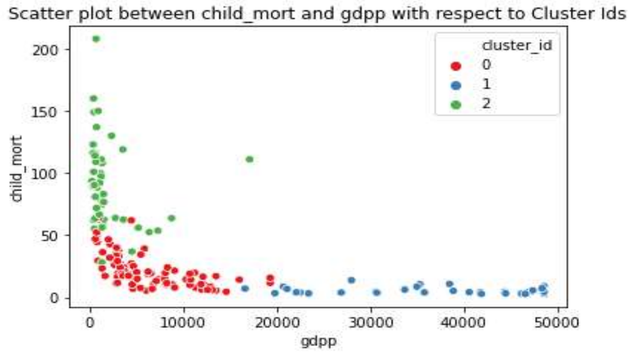


]:

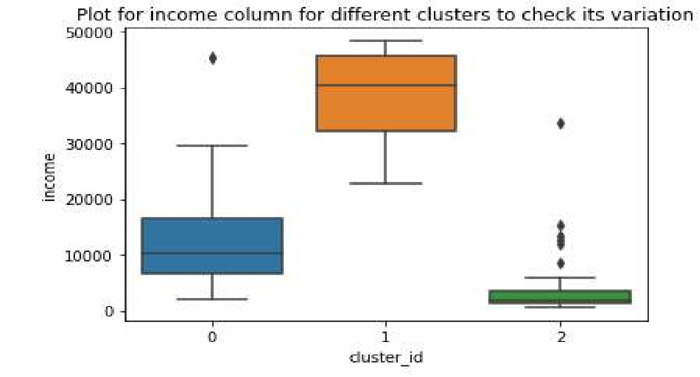
	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	cluster_id
0	Afghanistan	90.2	55.30	41.9174	248.297	1610.0	9.44	56.2	5.820	553	2
1	Albania	16.6	1145.20	267.8950	1987.740	9930.0	4.49	76.3	1.650	4090	0
2	Algeria	27.3	1712.64	185.9820	1400.440	12900.0	16.10	76.5	2.890	4460	0
3	Angola	119.0	2199.19	100.6050	1514.370	5900.0	20.87	60.1	5.861	3530	2
4	Antigua and Barbuda	10.3	5551.00	735.6600	7185.800	19100.0	1.44	76.8	2.130	12200	0



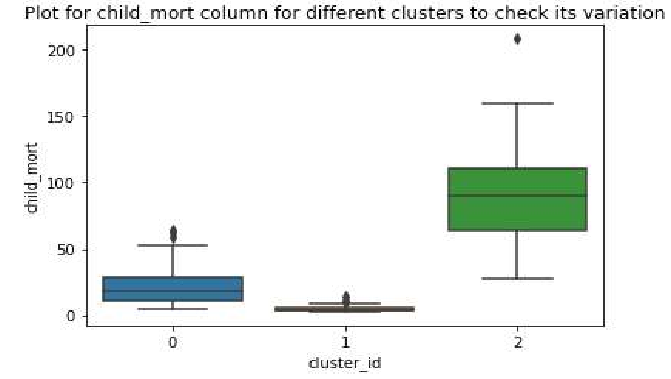
Scatter plot between income and gdpp with respect to Cluster Ids



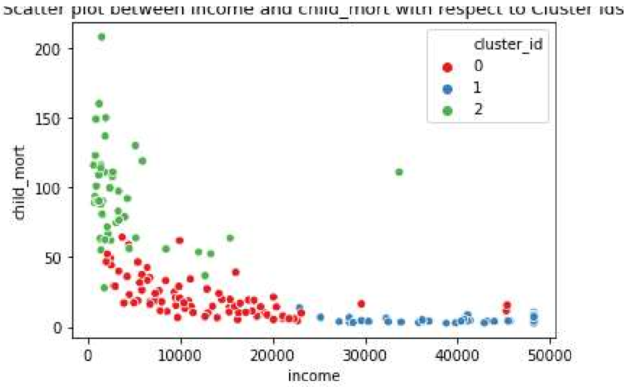
Scatter plot between child_mort and gdpp with respect to Cluster Ids



Plot for income column for different clusters to check its variation



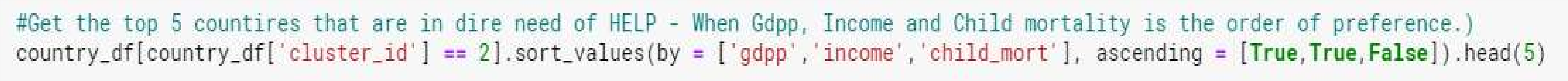
Plot for child_mort column for different clusters to check its variation



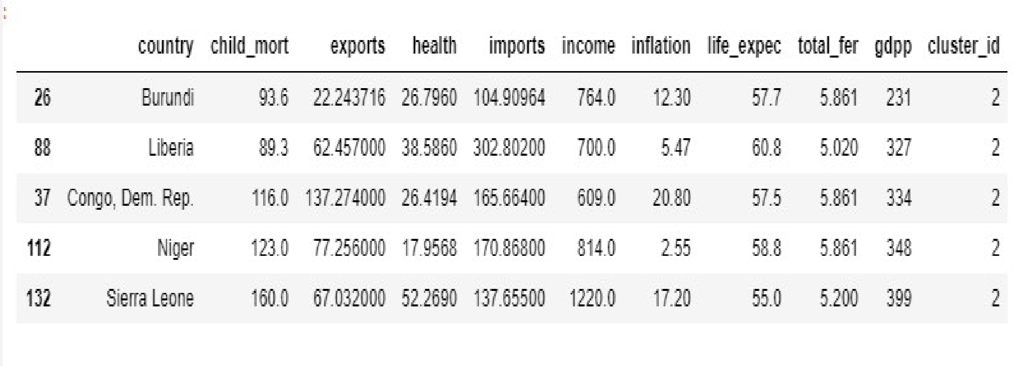
Scatter plot between income and child_mort with respect to Cluster Ids

CLUSTER PROFILING AND OBSERVATIONS

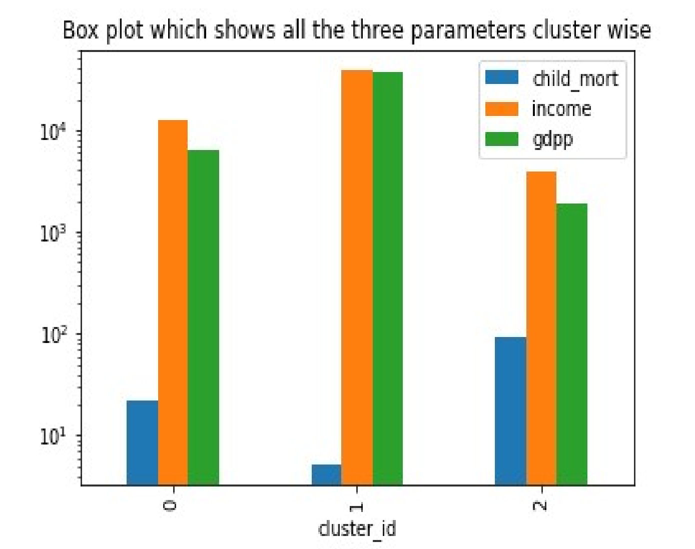
- Cluster 0: Medium child mortality rate, gdpp and income : Developing countries.
- Cluster 1: Less Child mortality rate, high gdpp and income : Developed countries.
- Cluster 2: High Child mortality rate, less gdpp and income : Under developed countries. Which we need to target.



```
#Get the top 5 countires that are in dire need of HELP - When Gdpp, Income and Child mortality is the order of preference.)
country_df[country_df['cluster_id'] == 2].sort_values(by = ['gdpp','income','child_mort'], ascending = [True,True,False]).head(5)
```



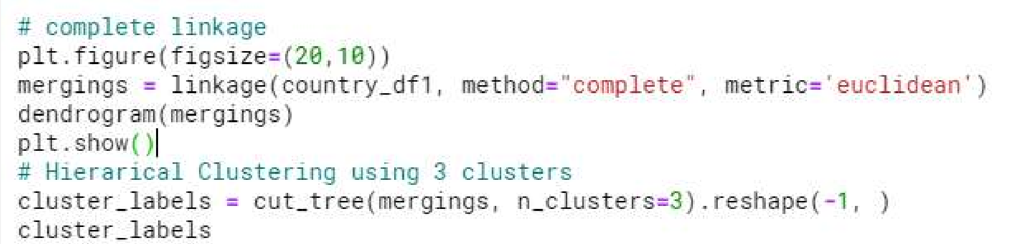
	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	cluster_id
26	Burundi	93.6	22.243716	26.7960	104.90964	764.0	12.30	57.7	5.861	231	2
88	Liberia	89.3	62.457000	38.5860	302.80200	700.0	5.47	60.8	5.020	327	2
37	Congo, Dem. Rep.	116.0	137.274000	26.4194	165.66400	609.0	20.80	57.5	5.861	334	2
112	Niger	123.0	77.256000	17.9568	170.86800	814.0	2.55	58.8	5.861	348	2
132	Sierra Leone	160.0	67.032000	52.2690	137.65500	1220.0	17.20	55.0	5.200	399	2



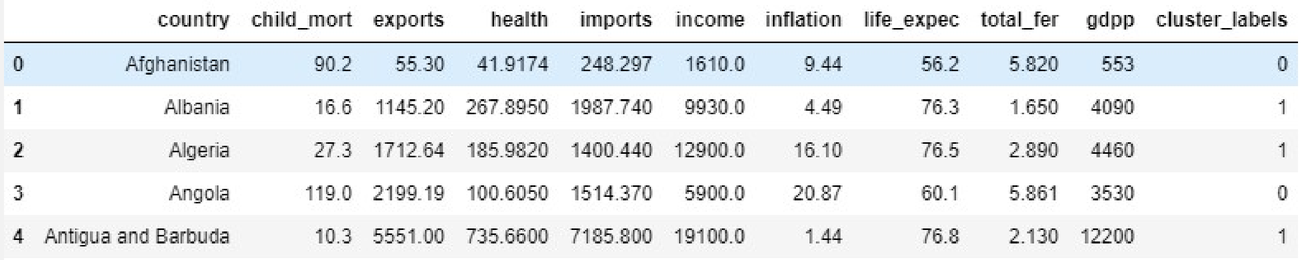
Box plot which shows all the three parameters cluster wise

DATA MODELING – HIERARCHICAL CLUSTERING

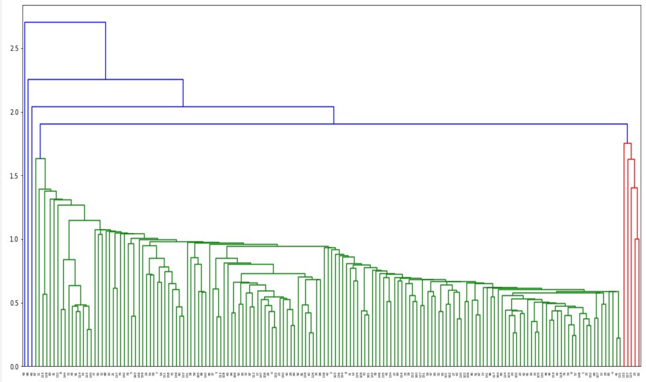
- K Value taken is 3 and have used both single and complete linkage and found complete linkage gives better insights.

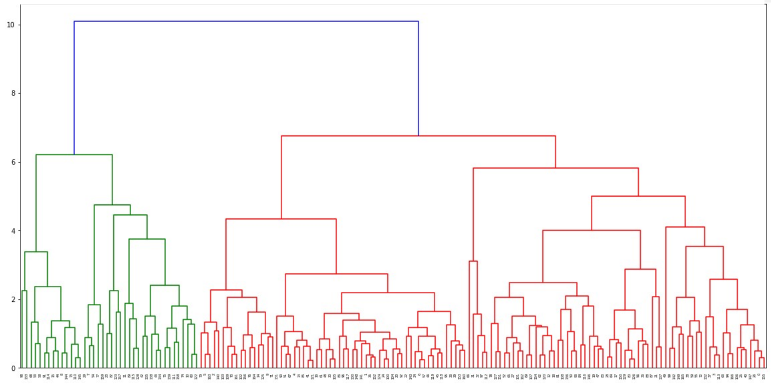


```
# complete linkage
plt.figure(figsize=(20,10))
mergings = linkage(country_df1, method="complete", metric='euclidean')
dendrogram(mergings)
plt.show()
# Hierarical Clustering using 3 clusters
cluster_labels = cut_tree(mergings, n_clusters=3).reshape(-1, )
cluster_labels
```

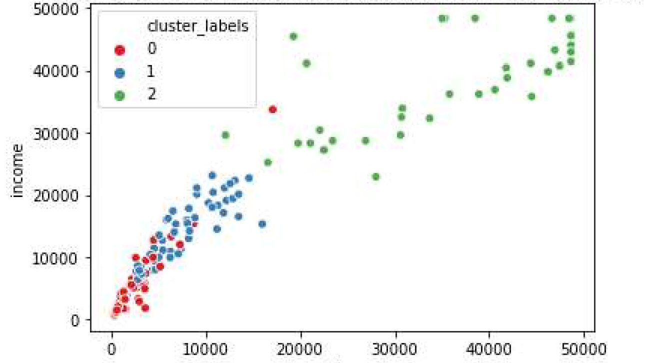


	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	cluster_labels
0	Afghanistan	90.2	55.30	41.9174	248.297	1610.0	9.44	56.2	5.820	553	0
1	Albania	16.6	1145.20	267.8950	1987.740	9930.0	4.49	76.3	1.650	4090	1
2	Algeria	27.3	1712.64	185.9820	1400.440	12900.0	16.10	76.5	2.890	4460	1
3	Angola	119.0	2199.19	100.6050	1514.370	5900.0	20.87	60.1	5.861	3530	0
4	Antigua and Barbuda	10.3	5551.00	735.6600	7185.800	19100.0	1.44	76.8	2.130	12200	1

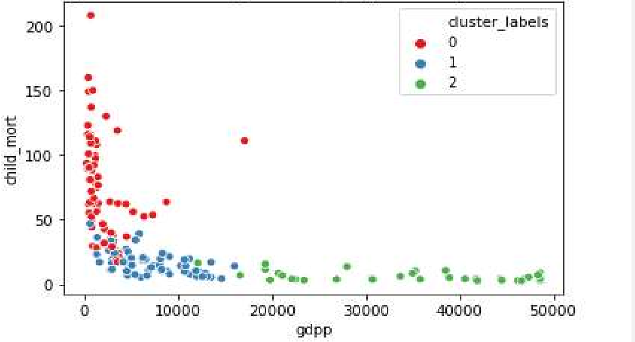




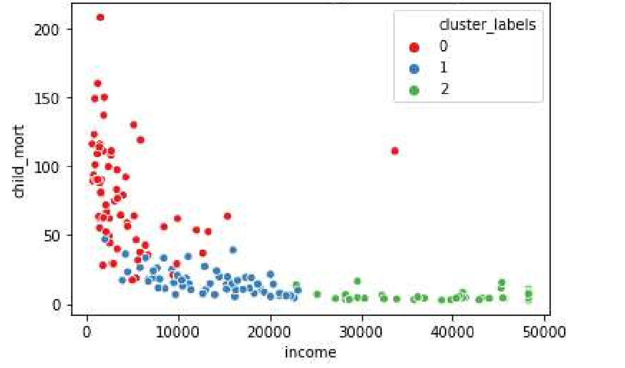
Scatter plot between income and gdpp with respect to Cluster Ids



Scatter plot between child_mort and gdpp with respect to Cluster Ids

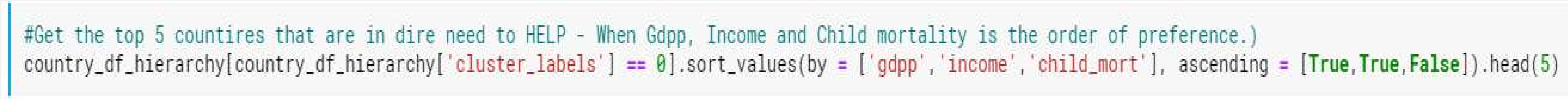


Scatter plot between income and child_mort with respect to Cluster Ids

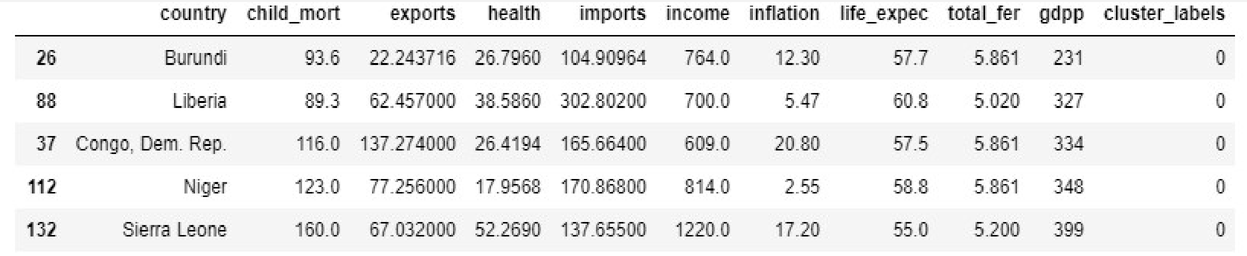


CLUSTER PROFILING AND OBSERVATIONS

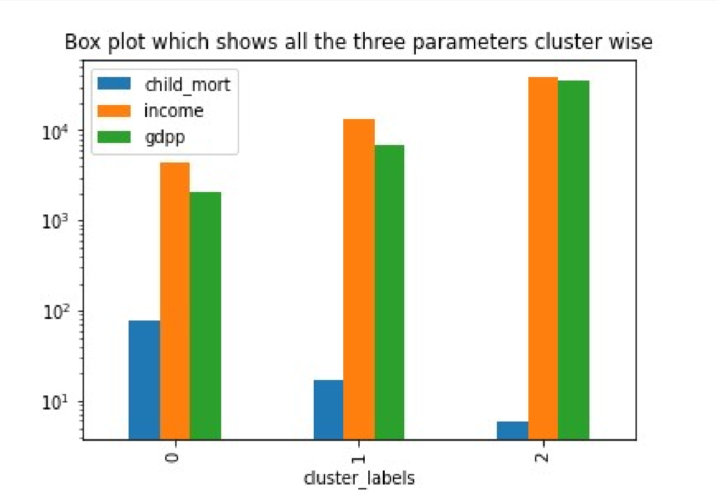
- Cluster 0: High Child mortality rate, less gdpp and income : Under developed countries. Which we need to target
- Cluster 1: Medium child mortality rate, gdpp and income : Developing countries.
- Cluster 2: Less Child mortality rate, high gdpp and income : Developed countries.



```
#Get the top 5 countires that are in dire need to HELP - When Gdpp, Income and Child mortality is the order of preference.)
country_df_hierarchy[country_df_hierarchy['cluster_labels'] == 0].sort_values(by = ['gdpp','income','child_mort'], ascending = [True,True,False]).head(5)
```



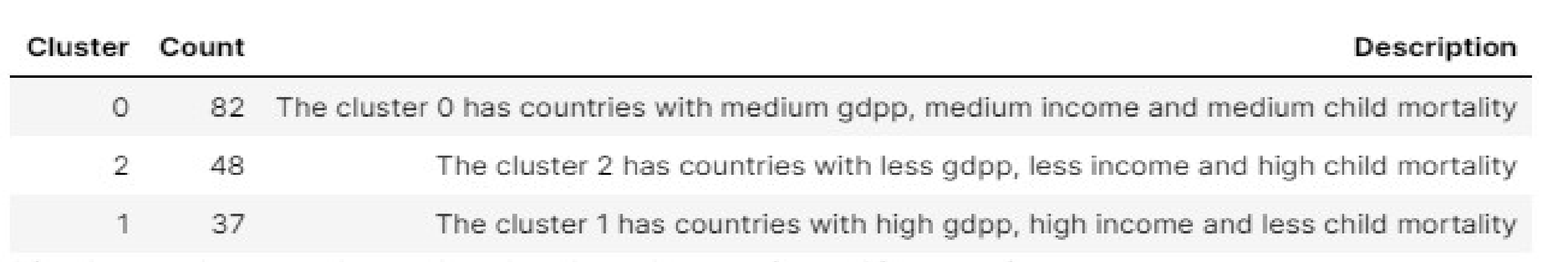
	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	cluster_labels
26	Burundi	93.6	22.243716	26.7960	104.90964	764.0	12.30	57.7	5.861	231	0
88	Liberia	89.3	62.457000	38.5860	302.80200	700.0	5.47	60.8	5.020	327	0
37	Congo, Dem. Rep.	116.0	137.274000	26.4194	165.66400	609.0	20.80	57.5	5.861	334	0
112	Niger	123.0	77.256000	17.9568	170.86800	814.0	2.55	58.8	5.861	348	0
132	Sierra Leone	160.0	67.032000	52.2690	137.65500	1220.0	17.20	55.0	5.200	399	0



Box plot which shows all the three parameters cluster wise

INFERENCES AND OBSERVATIONS

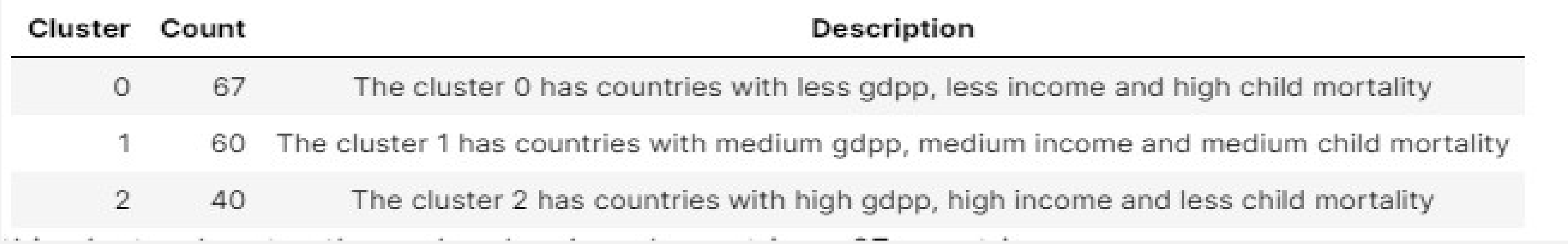
- The optimal value of k for clustering is 3
- K means Clustering model : The number of countries under each cluster are:



Cluster	Count	Description
0	82	The cluster 0 has countries with medium gdpp, medium income and medium child mortality
2	48	The cluster 2 has countries with less gdpp, less income and high child mortality
1	37	The cluster 1 has countries with high gdpp, high income and less child mortality

We need to concentrate on countries belonging to Cluster 2 as this cluster denotes the under developed countries.- 48 countries

- For Hierarchal Clustering model : The number of countries under each cluster are :



Cluster	Count	Description
0	67	The cluster 0 has countries with less gdpp, less income and high child mortality
1	60	The cluster 1 has countries with medium gdpp, medium income and medium child mortality
2	40	The cluster 2 has countries with high gdpp, high income and less child mortality

- We need to concentrate on countries belonging to Cluster 0 as this cluster denotes the under-developed countries - 67 countries

FINAL RECOMMENDATION

- The top five countries that need AID are (same countries obtained by k means and hierarchical algorithm)

1. Burundi,
2. Liberia
3. Congo. Dem. Rep
4. Niger
5. Sierra Leone

- I have considered the countries first with less gdpp followed by less income and high child mortality rate because
- 1. Gross Domestic Product of a country indicates the total value of production of goods and services of that country and thus indicates that the above five countries have the least production happening in their country. This in turn would lead to the need of requiring external support from other countries. When the financial help is provided to the countries based on the low GDP indicator, such countries should be able to produce more goods and services, which in turn helps in the countries overall development.
- 2. Low Income is a sign of the inability to purchase commodities or services which can keep a family comfortable or in dire need of resources in case of very low income families. As such, foreign help would provide the required help to such families at a lower cost.
- 3. Child mortality rate can be the third indicator as it is not directly linked to the resources available, as it is affected to other reasons like incomplete care of the mother during pregnancy, irresponsibility of the health care workers, etc. Thus, child moratlity rate can be considered as the third factor in this problem, keeping in mind the malnutrition and other financial problems the country is facing.