PCA AND CLUSTERING ASSIGNMENT

ANALYSIS AND VISUALIZATION

By :- Siddharth

PROCEDURE IN THE ASSIGNMENT

1. Data Understanding.
    a. Hint: Don't forget to read the data description properly.
2. Perform PCA.
    a. Data Standardization
    b. Perform PCA and choose the PCs that defines more than 85% variance.
    c. Run the PCA with the chosen number.
3. Perform Clustering.
    a. Data preparation for clustering.
        i. Outlier treatment
        ii. Hopkins check
    b. Clustering
        i. K-MEANS
            1. Run K-Means and choose K using both Elbow and Silhouette score
            2. Run K-Means with the chosen K
            3. Visualize the clusters
            4. Clustering profiling.

DATA UNDERSTANDING

- The data for this assignment was based on a comparison between countries on various parameters like child mortality rate, export, import, income, gdpp, health, inflation and life expectancy.
- Structure of dataframe after importing this csv was as follows:

  country 167 non-null object ,child_mort 167 non-null float64

  exports 167 non-null float64 ,health 167 non-null float64

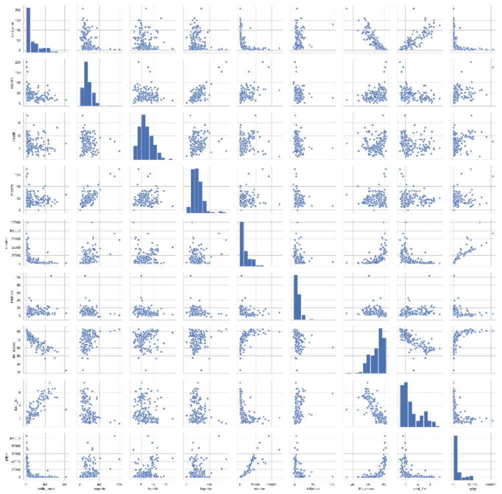
  imports 167 non-null float64 ,income 167 non-null int64

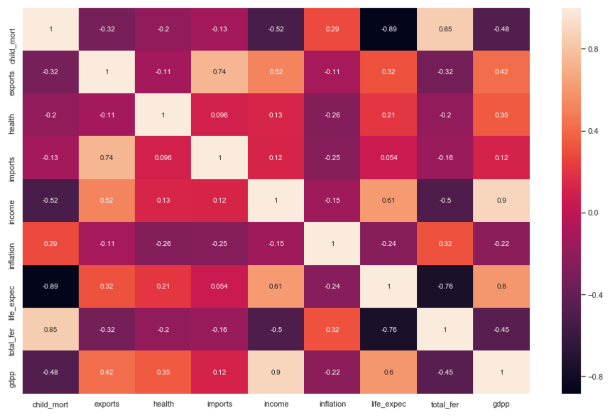
  inflation 167 non-null float64 ,life_expec 167 non-null float64

  total_fer 167 non-null float64 ,gdpp 167 non-null int64

DATA UNDERSTANDING

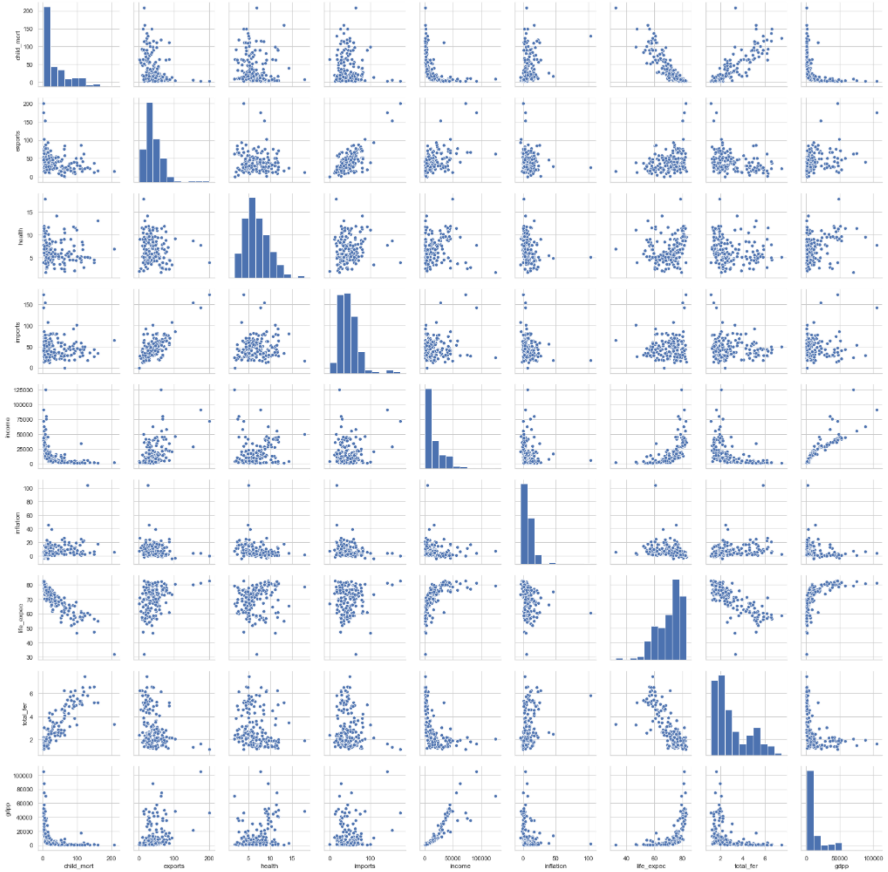
- Null value check
- Set ‘country’ as index.
- Percentile check
- Multiple Bivariate analysis using pairplot.
- Correlation heatmap
- Check for Outliers.



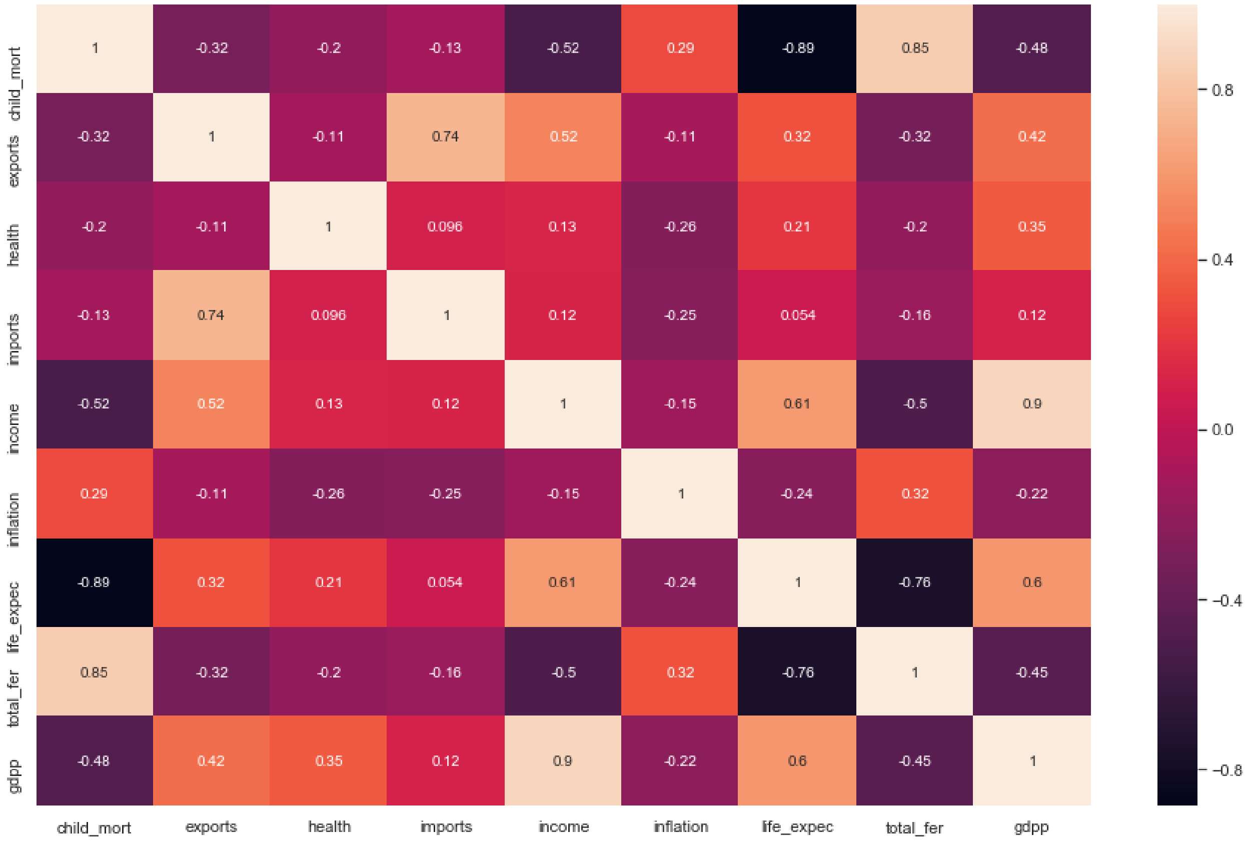


DATA UNDERSTANDING

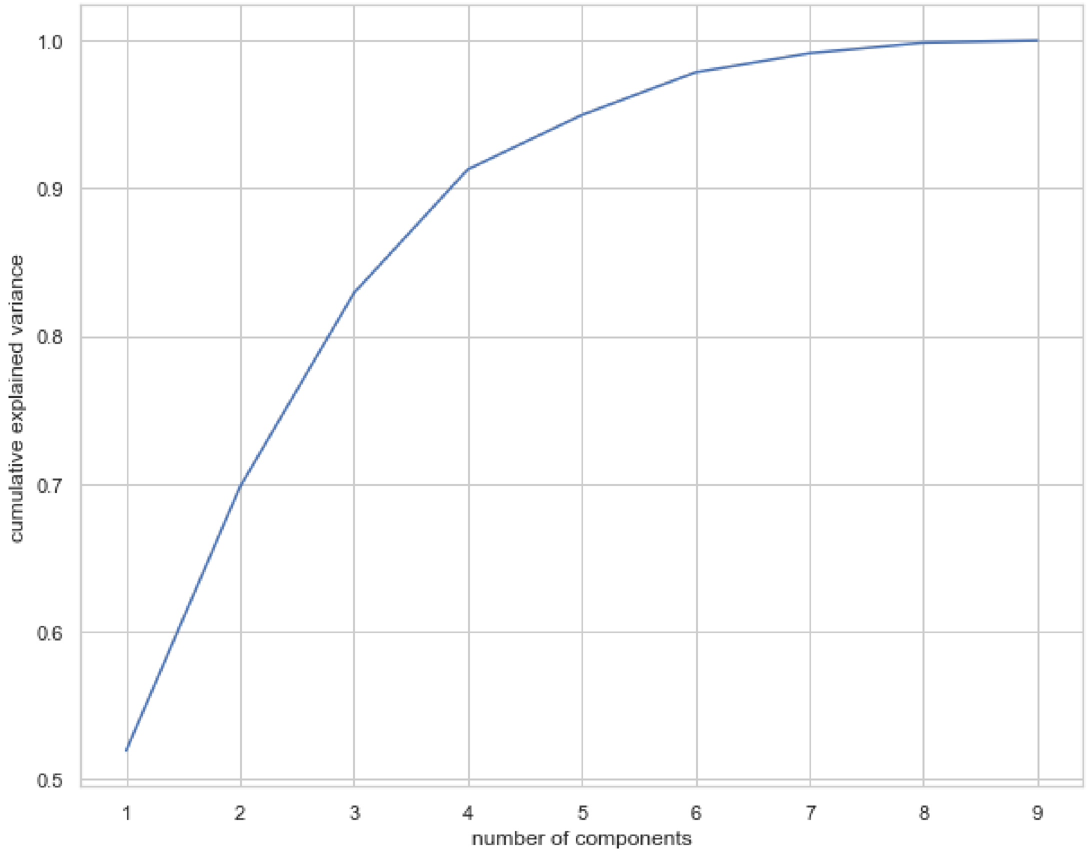
Pairplot



Correlation heatmap



PRINCIPAL CLUSTERING ANALYSIS (DIMENSIONALITY REDUCTION)



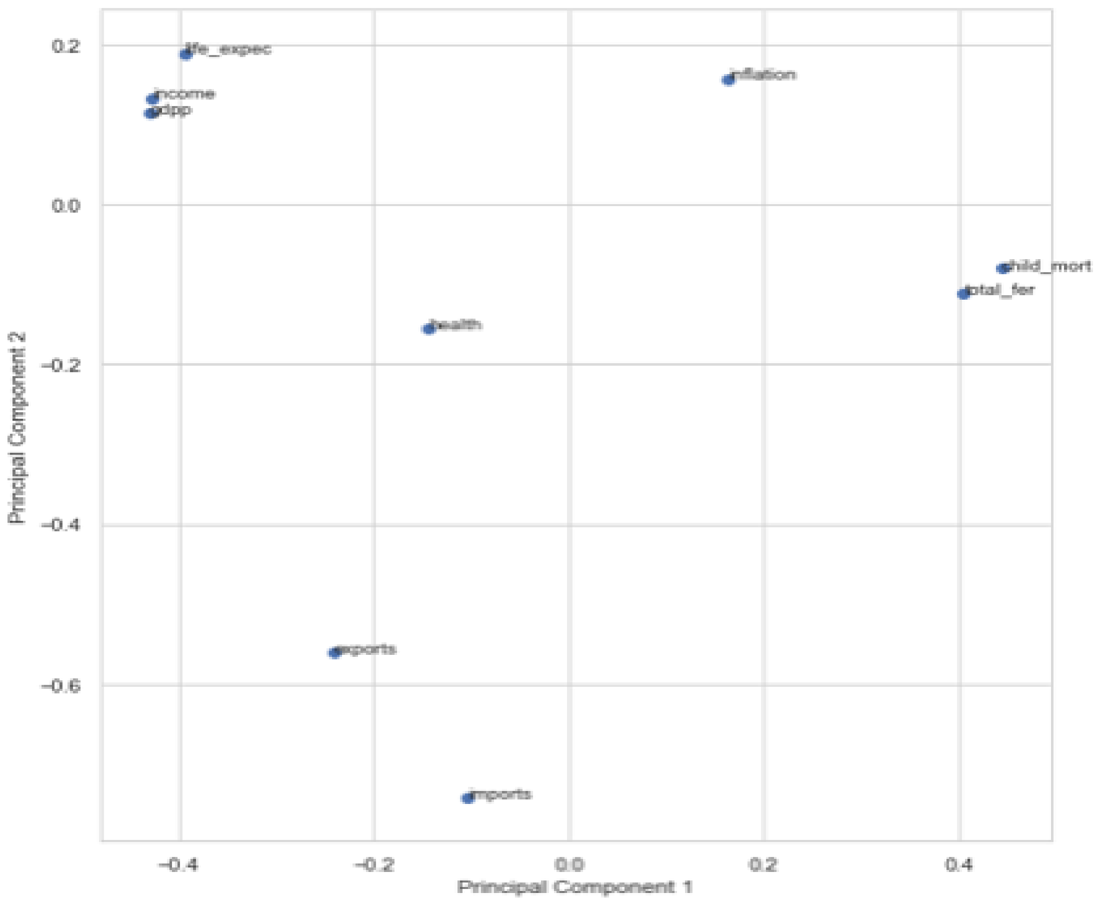
We take 5 cluster at pca(0.94), there is no point of taking 6 clusters at pca(0.96).

And have named these clusters as :

- PC1, PC2, PC3, PC4 AND PC5.

We performed outlier analysis after performing pca and now have 164 countries out of 166(initially).

PRINCIPAL CLUSTERING ANALYSIS (DIMENSIONALITY REDUCTION)



Visualization of principal components

CHECK HOPKINS STATISTICS

Hopkins Statistics:¶

- The Hopkins statistic, is a statistic which gives a value which indicates the cluster tendency, in other words: how well the data can be clustered.
- If the value is between {0.01, ...,0.3}, the data is regularly spaced.
- If the value is around 0.5, it is random.
- If the value is between {0.7, ..., 0.99}, it has a high tendency to cluster.

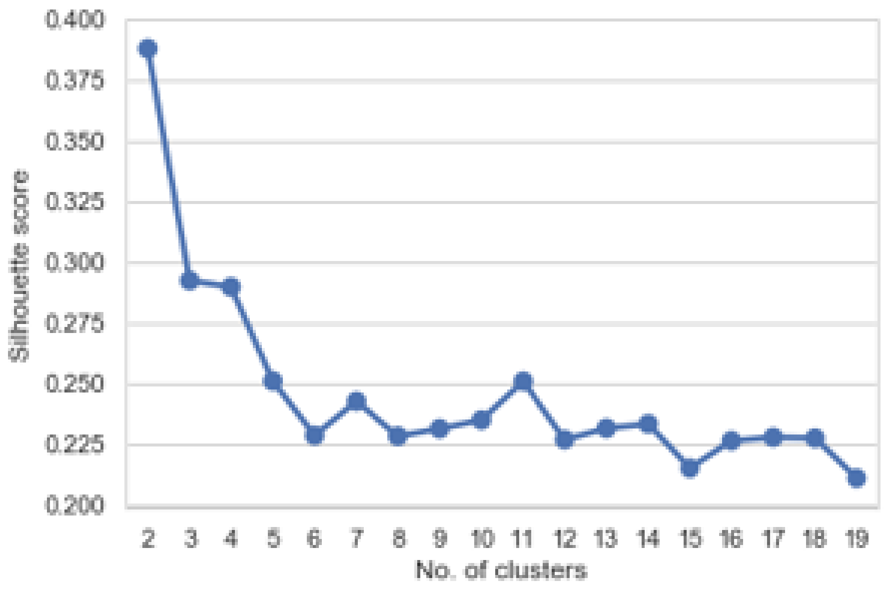
```
print("DF_PCA: ", hopkins(df_pca))
print("DF_scaled: ", hopkins(df_scaled))
```

```
DF_PCA:  0.7147223777638881
DF_scaled:  0.7421177522783486
```

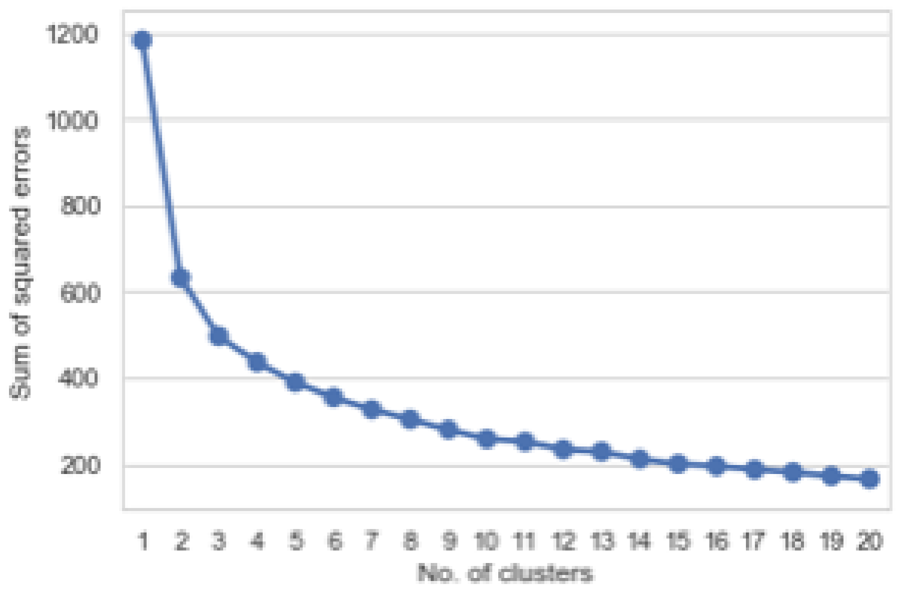
CLUSTERING

- **Select optimum number of clusters using silhouette score, no of squared errors.**

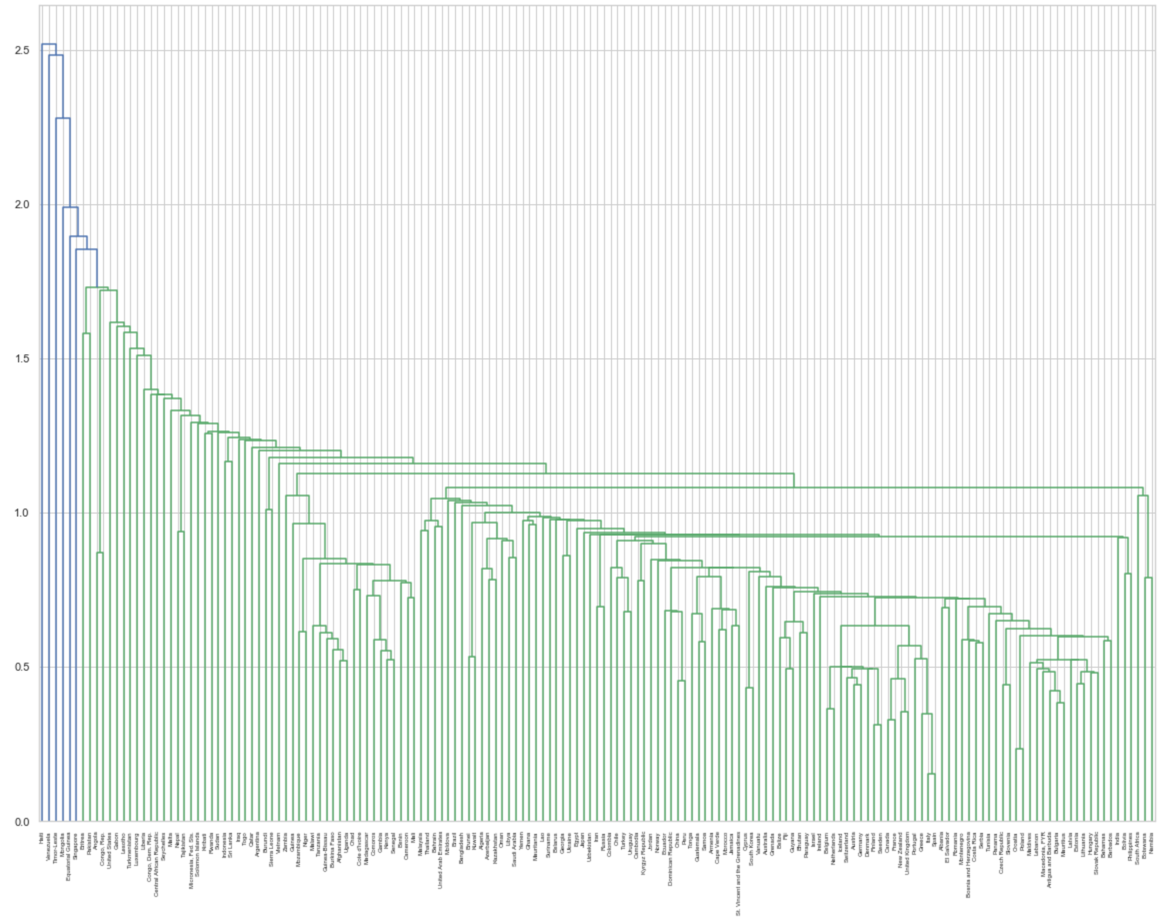
Text(0, 0.5, 'Silhouette score')



Text(0, 0.5, 'Sum of squared errors')



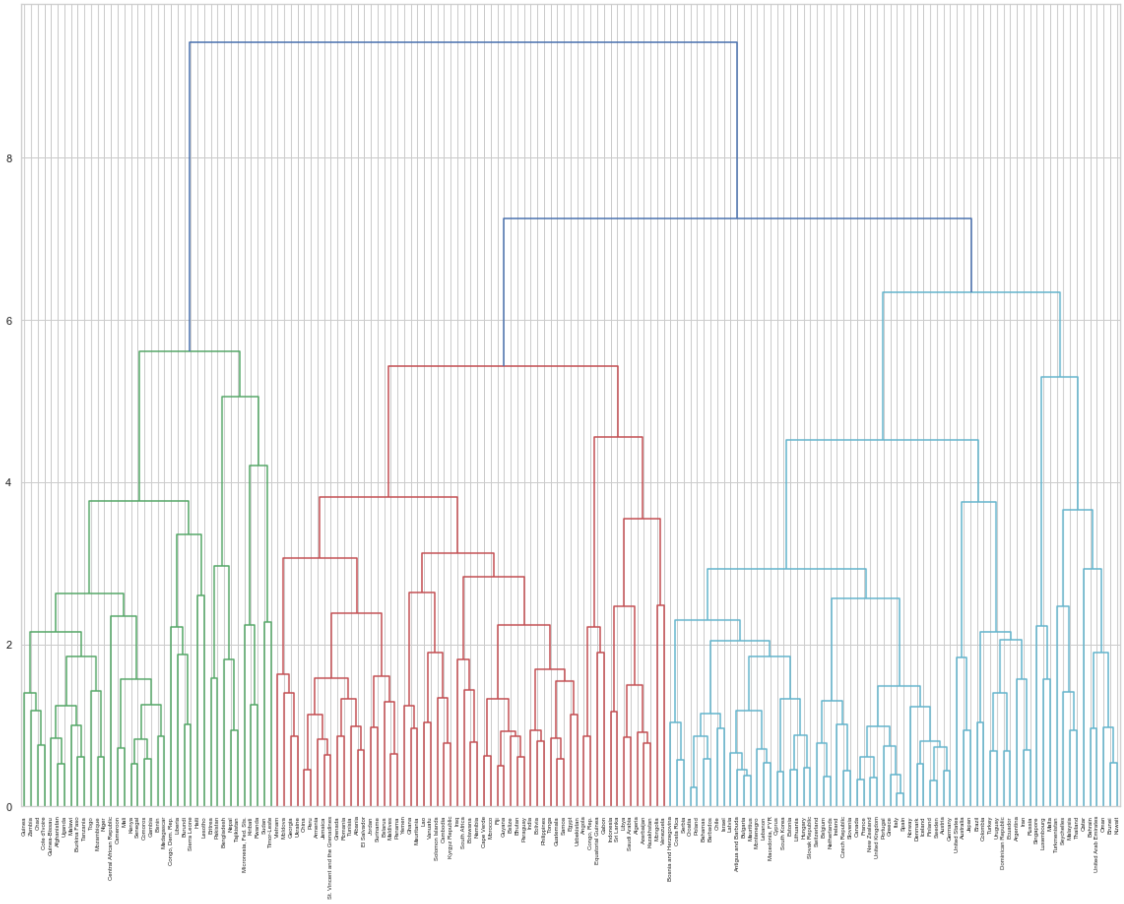
OPTIMUM NUMBER OF CLUSTERS THROUGH HEIRARCHICAL CLUSTERING




```python
from scipy.cluster.hierarchy import linkage
from scipy.cluster.hierarchy import dendrogram
from scipy.cluster.hierarchy import cut_tree
```

```python
plt.figure(figsize=(20,15))
mergings_s = linkage(df_pca, method = "single", metric='euclidean')
dendrogram(mergings_s, labels=df_pca.index, leaf_rotation=90, leaf_font_size=6)
plt.show()
```

OPTIMUM NUMBER OF CLUSTERS THROUGH HEIRARCHICAL CLUSTERING




```
plt.figure(figsize=(20,15))
mergings_c = linkage(df_pca, method = "complete", metric='euclidean')
dendrogram(mergings_c, labels=df_pca.index, leaf_rotation=90, leaf_font_size=6)
plt.show()
```

We can take the number of clusters to be either 3 or 4

K-MEANS

```
# Kmeans with K=4
K_means_4 = KMeans(n_clusters = 4, max_iter=50)
K_means_4.fit(df_pca)
```

```
KMeans(algorithm='auto', copy_x=True, init='k-means++', max_iter=50,
       n_clusters=4, n_init=10, n_jobs=None, precompute_distances='auto',
       random_state=None, tol=0.0001, verbose=0)
```

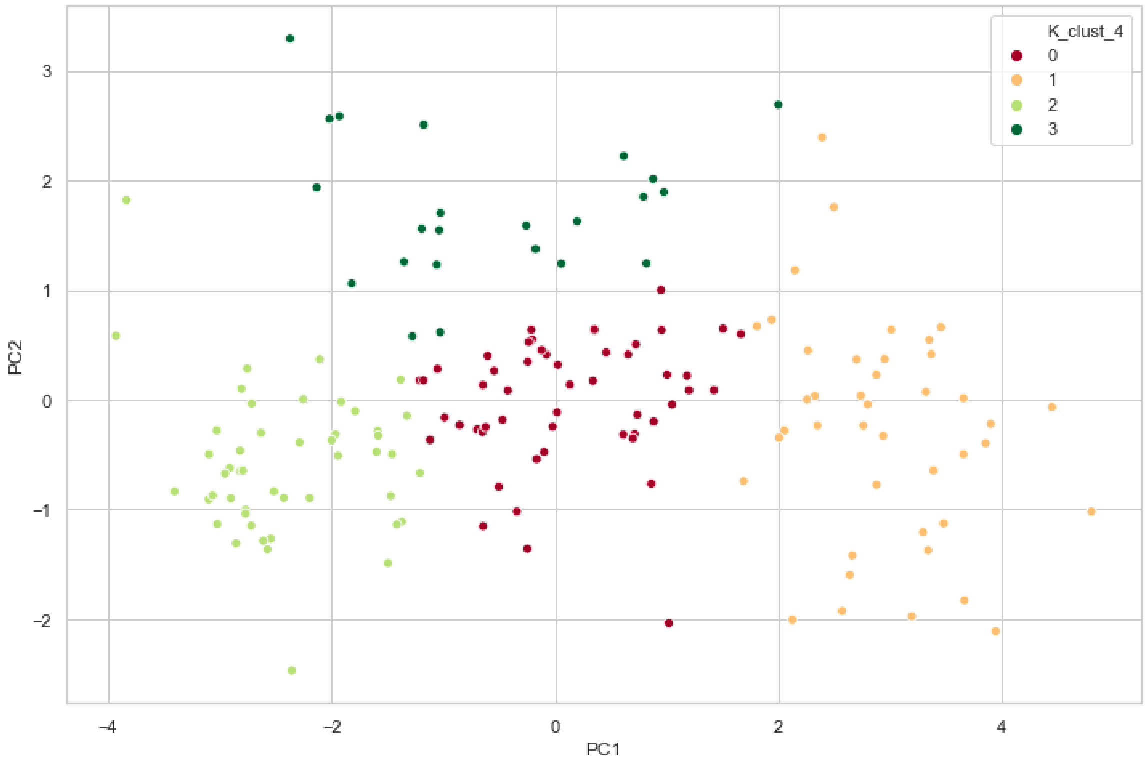
K_clust_4	0	1	2	3
H_clust_4				
0	2	36	0	0
1	41	6	0	12
2	9	0	45	1
3	0	0	3	10

Effectively, K-means clustering has broken down the cluster '1' of Hierarchical clustering - into 2 sub-clusters

We will use the clusters formed by K-means for further analysis

Combining original data, principal components, K-means cluster IDs & Hierarchical clustering cluster IDs

K-MEANS



Scatter Plot of 4 clusters,

We can clearly see that all 4 clusters a distinctly scattered across the space.

K-MEANS

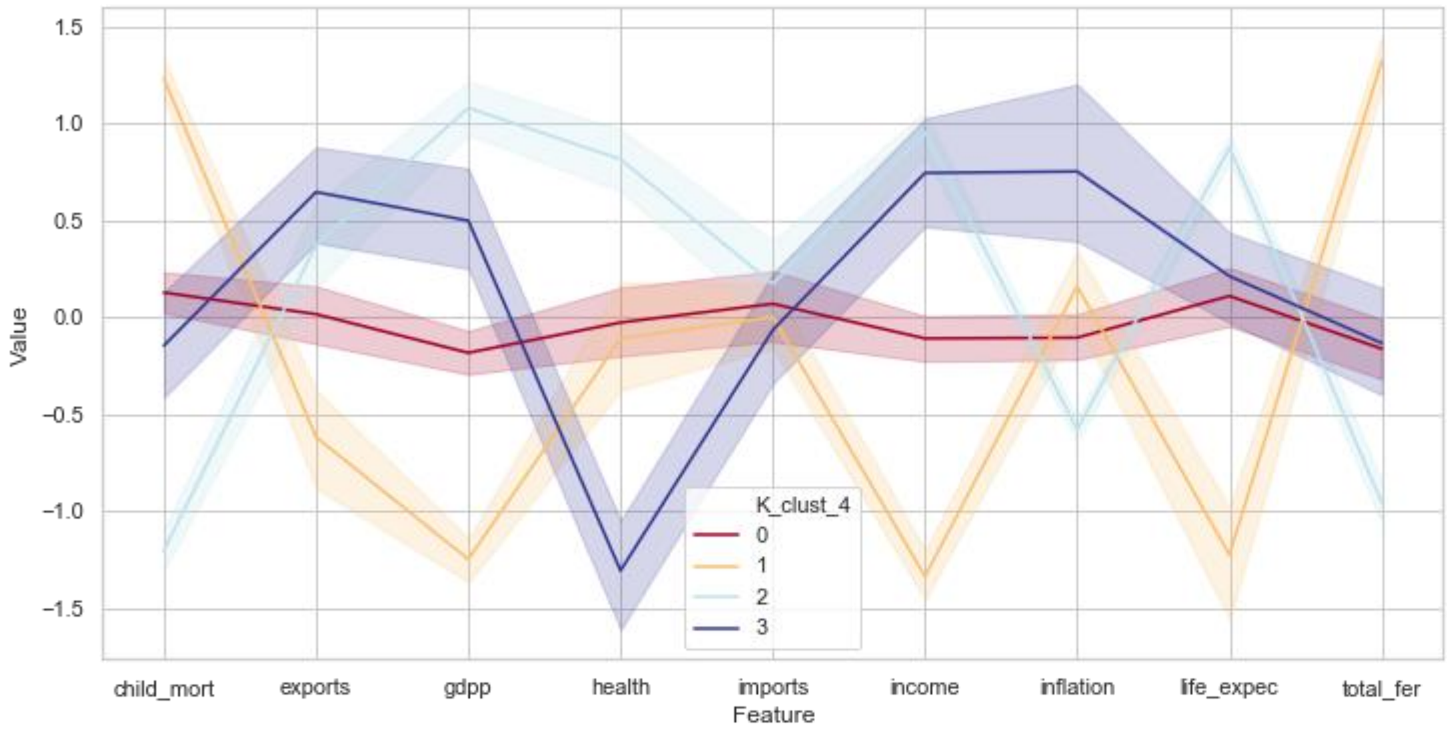
country	child_mort	exports	health	imports	income	life_expec	total_fer	gdpp	inflation	K_clust_4
Afghanistan	1.257285	-1.566676	0.450917	0.146381	-1.413059	-1.580003	1.667924	-1.460560	0.157336	1
Albania	-0.206196	-0.224045	0.105222	0.262136	0.071749	0.630275	-0.962078	-0.122592	-0.312347	0
Algeria	0.223939	0.187829	-0.963592	-0.376408	0.285304	0.649198	0.207329	-0.064683	0.789274	3
Angola	1.496866	0.818843	-1.864462	0.079772	-0.353135	-1.094994	1.786385	-0.221051	1.387054	1
Antigua and Barbuda	-0.618844	0.409060	-0.090571	0.543123	0.605603	0.677491	-0.429314	0.608191	-0.601749	2

Result we get after clustering.

	country	K_clust_4	Feature	Value
0	Afghanistan	1	child_mort	1.257285
1	Albania	0	child_mort	-0.206196
2	Algeria	3	child_mort	0.223939
3	Angola	1	child_mort	1.496866
4	Antigua and Barbuda	2	child_mort	-0.618844

We have taken a smaller dataframe with K_clust_4 and Feature with the value

LINEPLOT ANALYSIS



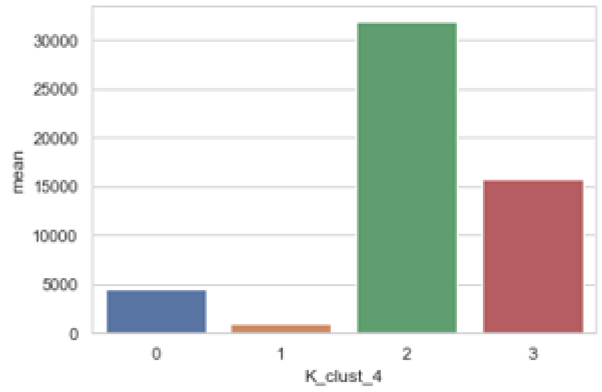
Observation: Cluster #1(in the above plot represented by the red line) contains countries that are in direct need of financial aid, since:

It has disproportionately high child mortality rate, total_fer & inflation.
It has lowest gdpp, income & life_expectancy.

BARPLOT FOR FEATURES VS CLUSTERS

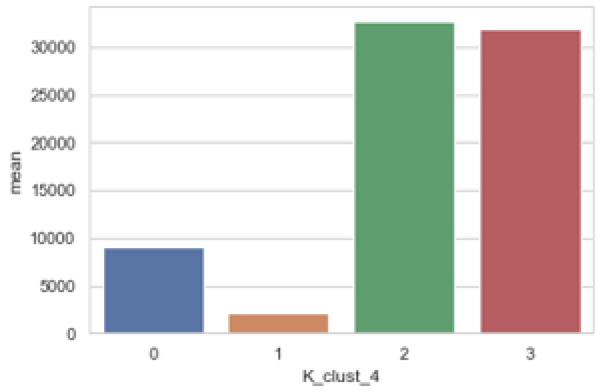
```
sns.barplot(x = cluster_summary.reset_index().K_clust_4, y = cluster_summary['gdpp']['mean'])
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x17fbfdfb6a0>
```




```
sns.barplot(x = cluster_summary.reset_index().K_clust_4, y = cluster_summary['income']['mean'])
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x17fc0f81550>
```

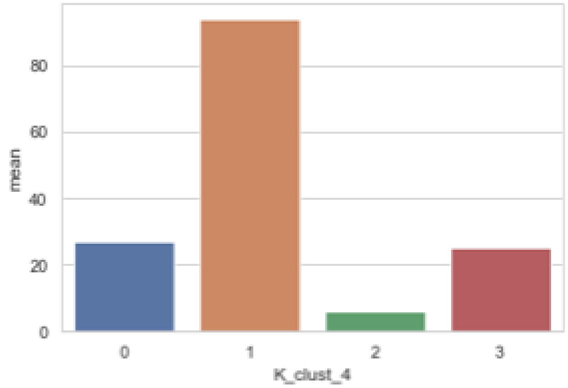


We can infer from these barplot that cluster #1 is really suffering.

BARPLOT FOR FEATURES VS CLUSTERS

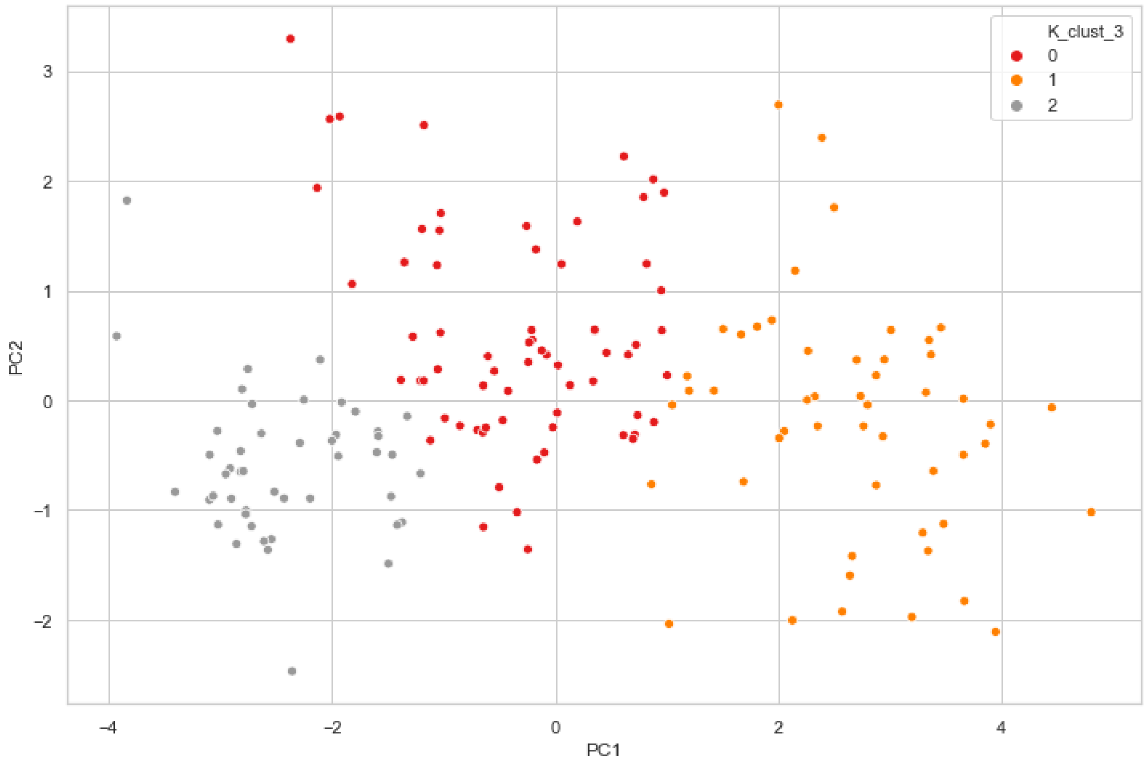
```
sns.barplot(x = cluster_summary.reset_index().K_clust_4, y = cluster_summary['child_mort']['mean'])
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x17fc0fe5b38>
```



We can infer from these barplot that cluster #1 is really suffering.

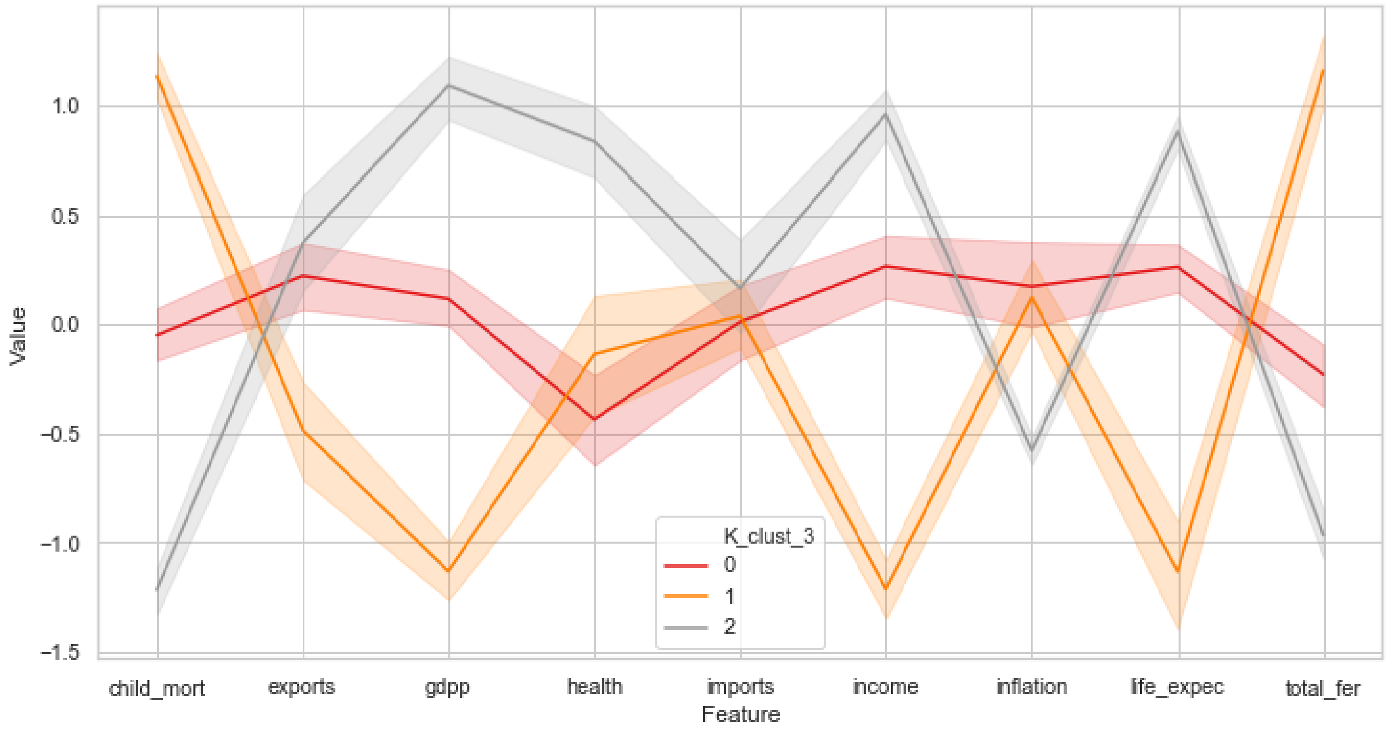
CLUSTER ANALYSIS FOR # OF CLUSTERS = 3



All 3 clusters are scattered distinctly in their own territory

Observation: Cluster #1 of 4-cluster K-means is the same as cluster #1 for 3-cluster K-means; This is the cluster with countries in dire need of financial aid.

CLUSTER ANALYSIS FOR # OF CLUSTERS = 3



Observation: Cluster #1(in the above plot represented by the red line) contains countries that are in direct need of financial aid, since:¶

It has disproportionately high child mortality rate, total_fer & inflation.
It has lowest gdpp, income & life_expectancy.

CONCLUSION

Child Mortality

	country	K_clust_4	Feature	Value
66	Haiti	1	child_mort	1.979683
130	Sierra Leone	1	child_mort	1.752837
32	Chad	1	child_mort	1.697036
31	Central African Republic	1	child_mort	1.691253
97	Mali	1	child_mort	1.618654
111	Niger	1	child_mort	1.525451
3	Angola	1	child_mort	1.496866
25	Burkina Faso	1	child_mort	1.474789
37	Congo, Dem. Rep.	1	child_mort	1.474789
64	Guinea-Bissau	1	child_mort	1.459752
17	Benin	1	child_mort	1.436694
40	Cote d'Ivoire	1	child_mort	1.436694
63	Guinea	1	child_mort	1.420973
28	Cameroon	1	child_mort	1.413004
106	Mozambique	1	child_mort	1.355066

Income

	country	K_clust_4	Feature	Value
697	Congo, Dem. Rep.	1	income	-2.206478
748	Liberia	1	income	-2.092822
686	Burundi	1	income	-2.021421
771	Niger	1	income	-1.969684
691	Central African Republic	1	income	-1.898672
766	Mozambique	1	income	-1.871555
754	Malawi	1	income	-1.777605
723	Guinea	1	income	-1.659760
808	Togo	1	income	-1.646157
790	Sierra Leone	1	income	-1.639440
784	Rwanda	1	income	-1.556804
724	Guinea-Bissau	1	income	-1.532974
753	Madagascar	1	income	-1.532974
696	Comoros	1	income	-1.521314
710	Eritrea	1	income	-1.515547

Gdpp

	country	K_clust_4	Feature	Value
1181	Burundi	1	gdpp	-2.044268
1243	Liberia	1	gdpp	-1.811877
1192	Congo, Dem. Rep.	1	gdpp	-1.797714
1266	Niger	1	gdpp	-1.770258
1285	Sierra Leone	1	gdpp	-1.678811
1248	Madagascar	1	gdpp	-1.655752
1261	Mozambique	1	gdpp	-1.646107
1186	Central African Republic	1	gdpp	-1.604350
1249	Malawi	1	gdpp	-1.585138
1205	Eritrea	1	gdpp	-1.552444
1303	Togo	1	gdpp	-1.544172
1219	Guinea-Bissau	1	gdpp	-1.467855
1155	Afghanistan	1	gdpp	-1.460560
1211	Gambia	1	gdpp	-1.449765
1279	Rwanda	1	gdpp	-1.448576

We can clearly see that countries are common in each of these dataframe with respect to (child_mort, income and gdpp). Some of those countries are:

Congo, Dem. Rep.
Sierra Leone
Niger
Mozambique
Guinea-Bissau
Central African Republic
Liberia

CONCLUSION

- Congo, Dem. Rep., Sierra Leone, Niger, Mozambique, Guinea-Bissau, Central African Republic and Liberia are the countries in dire need of financial aid as these countries are common among all three features.