Clustering Assignment

Submitted By:

Minakshi Maurya

Outline

- Problem Statement
- Data Understanding/Data Visualization
- Outlier Treatment
- Clustering
- Clustering Visualization
- Cluster Profiling
- Conclusion statement

Problem Statement

IDEAL:

CEO of HELP International NGO would be able to decide how to use raised money strategically and effectively by getting the countries who are in dire need of aid.

REALITY:

In reality we are having data for 167 companies with factors affecting it's development. From the raw dataset it will be difficult to decide that which courtiers are in dire need of aid.

CONSEQUENCES:

May be raised fund will go to the countries not in dire need of aid and countries in dire need of aid will get skip.

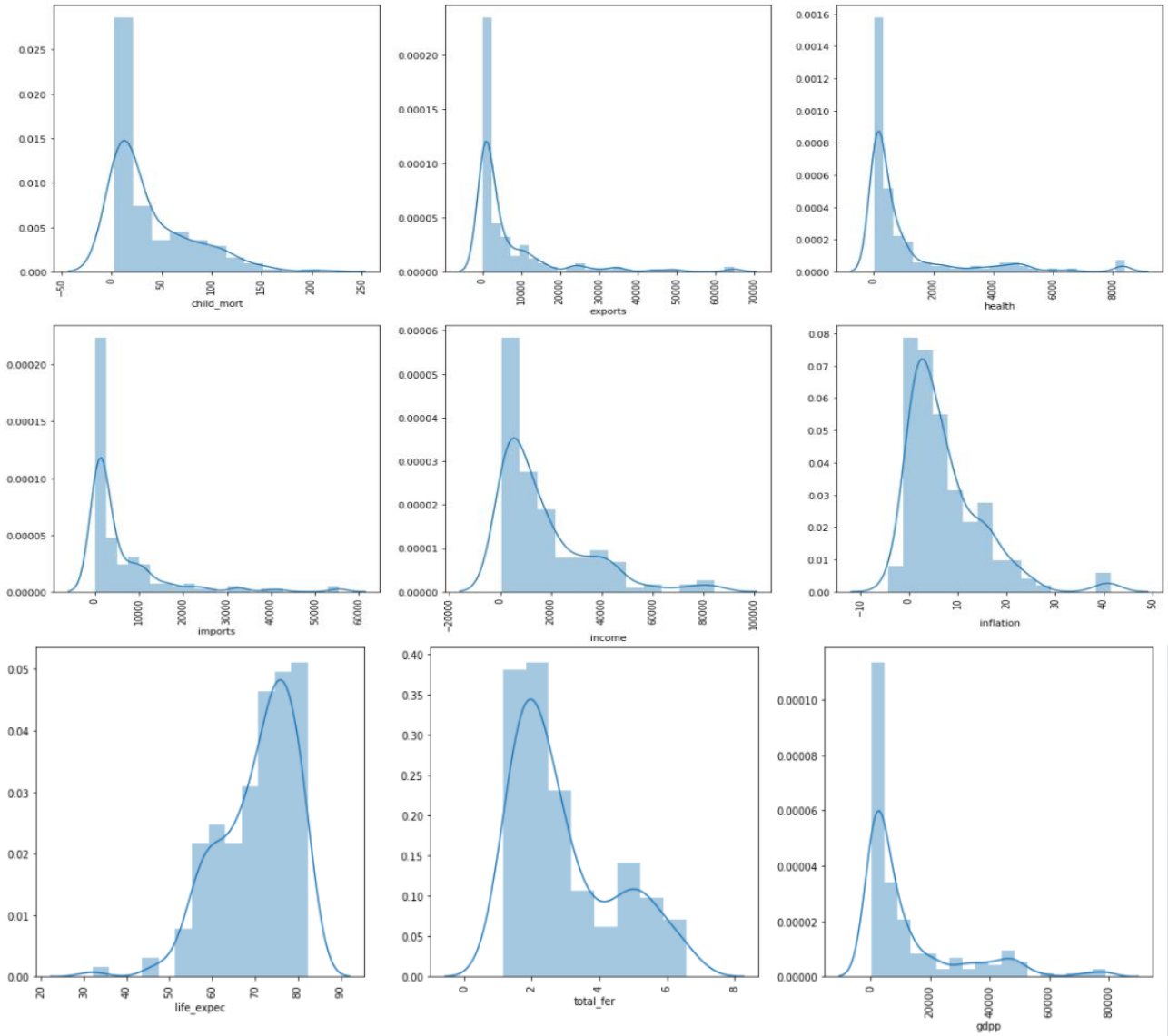
PROPOSAL:

Clustering will be done to categorize the companies in different cluster based on the country development factors(Variables)

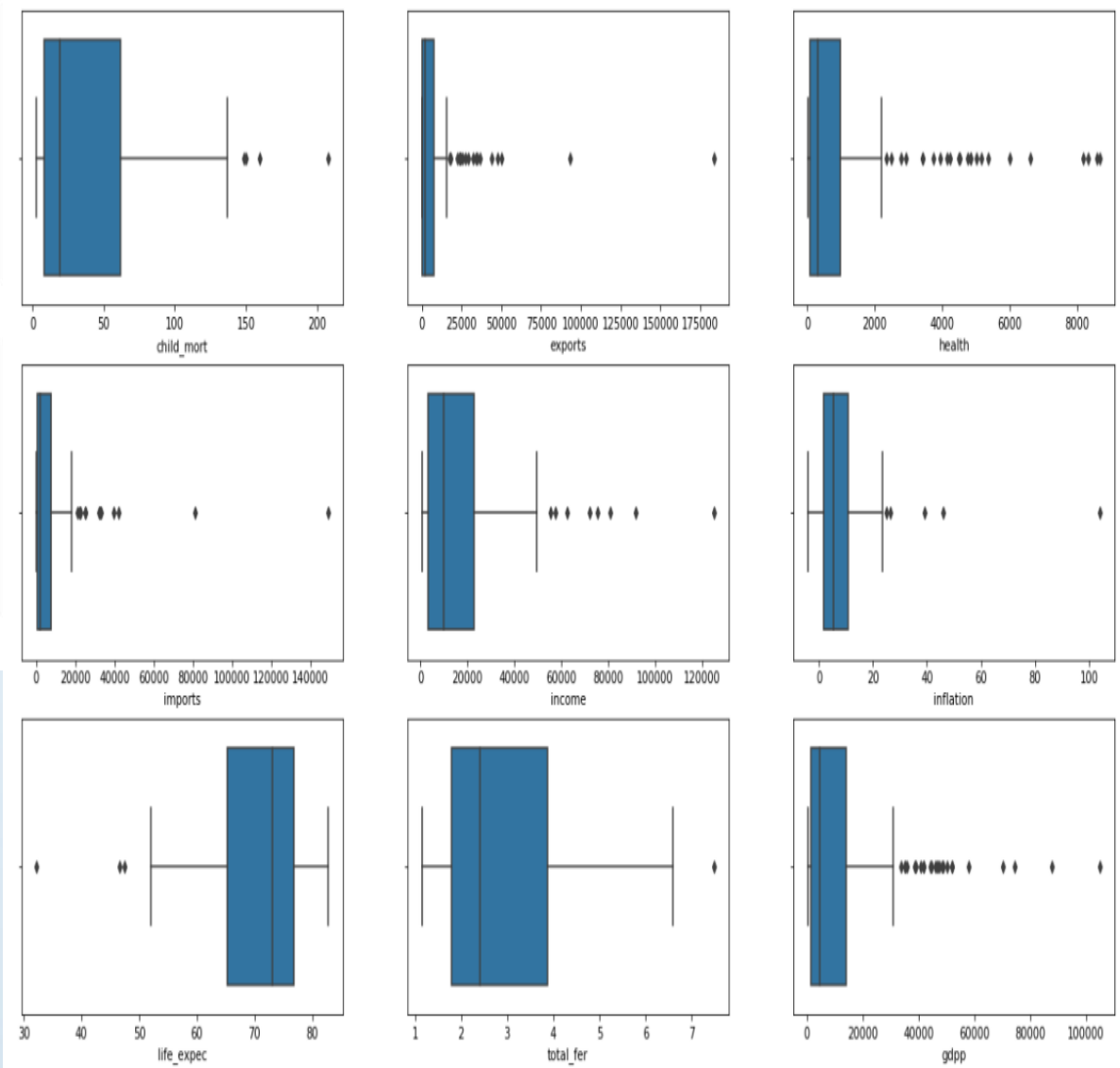
Data Understanding/Data Visualization

- Data set is having information for 167 companies with 9 country's development factor
- There is no missing values in Data set.
- Exports, Imports and Health are presented as percentage of GDP per capita, which needs to be converted to actual values to get correct inference.
- Continuous variables are distributed normally and having outliers
- Child_mort is inversely related to health, income, gdpp, life_expec, exports, imports. If these factors increases then child_mort decreases
- Imports, Exports ,GDPP, Health , Income and life_expec are positively corelated to each other.
- Child_mort, total_fer and inflation are positively corelated to each other

Data Understanding/Data Visualization continue…..

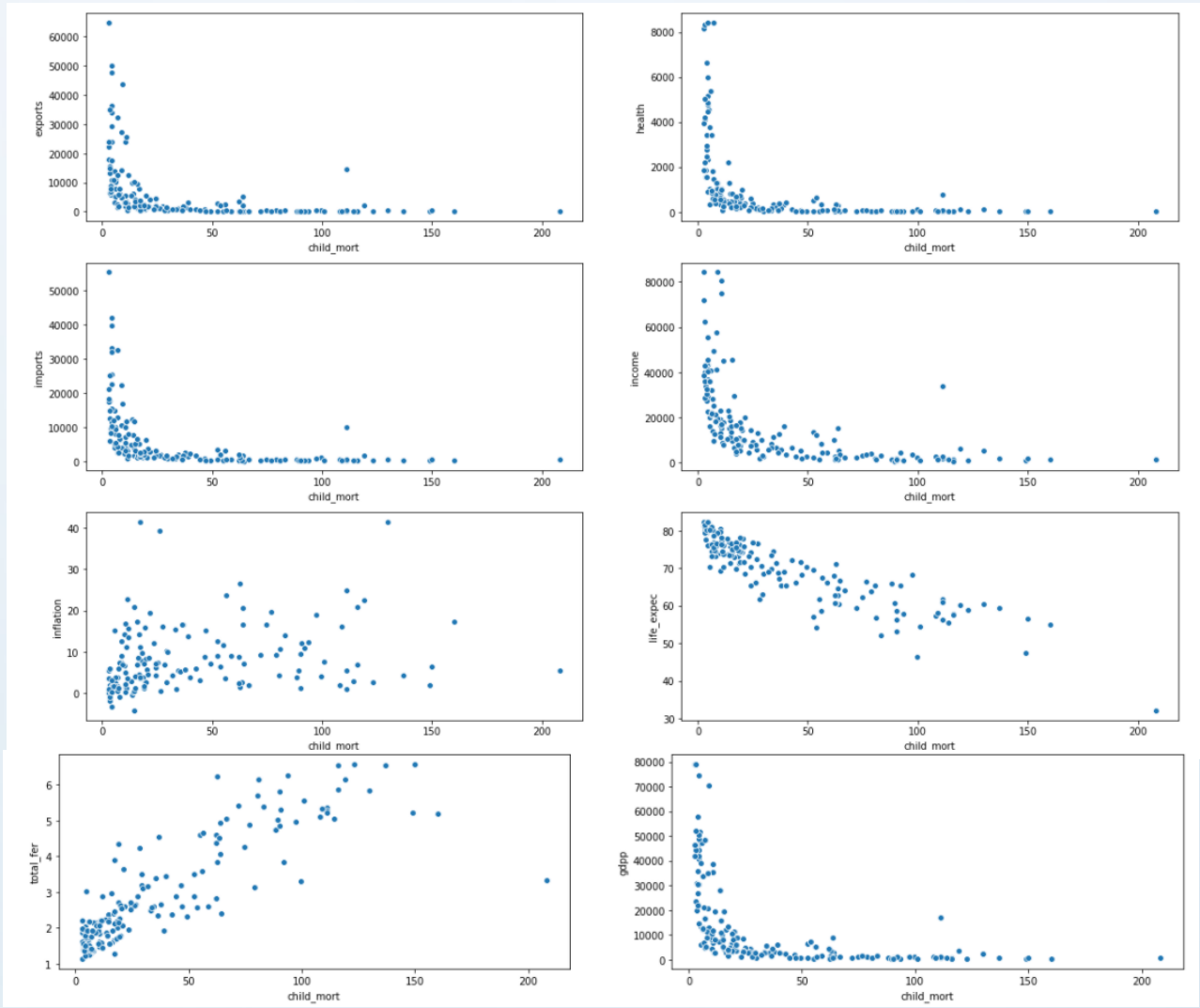


Variable Distribution

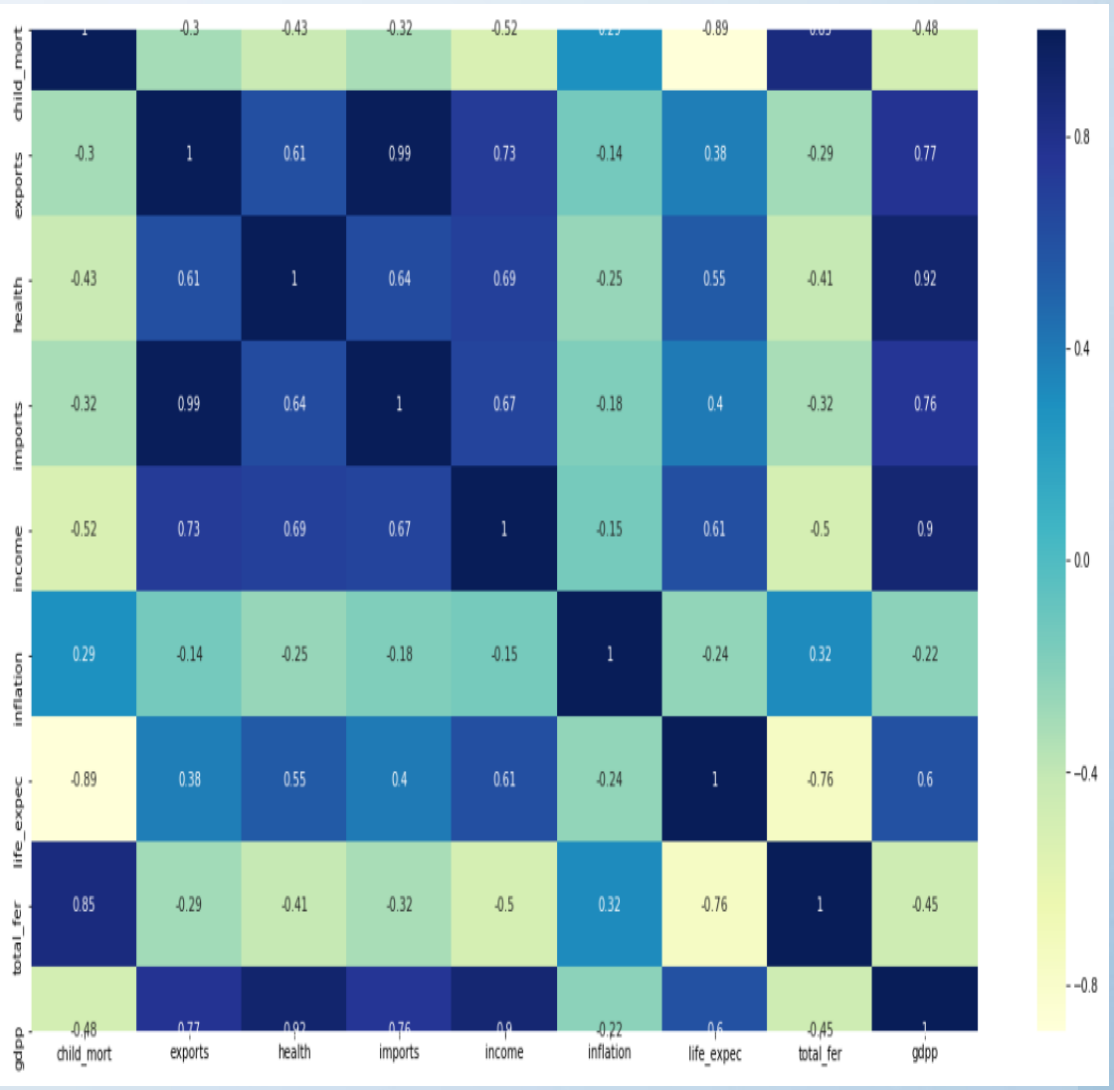


Outliers

Data Understanding/Data Visualization continue…..



Child mortality Vs other Variables



Correlation Matrix

Outlier Treatment

For Outlier Treatment soft caping has been used

- For Child_mort lower outlier has been treated and higher has been kept as it is, may be higher outliers companied need aid
- For other Variables higher outliers has been treated and lower outliers has been kept as it is.

```
q1 = df['child_mort'].quantile(0.01)
df['child_mort'][df['child_mort']<= q1] = q1

# Caping highier outlier for other variables
q3_exports = df['exports'].quantile(0.99)
df['exports'][df['exports']>= q3_exports] = q3_exports

q3_imports = df['imports'].quantile(0.99)
df['imports'][df['imports']>= q3_imports] = q3_imports

q3_health = df['health'].quantile(0.99)
df['health'][df['health']>= q3_health] = q3_health

q3_gdpp = df['gdpp'].quantile(0.99)
df['gdpp'][df['gdpp']>= q3_gdpp] = q3_gdpp

q3_life_expec = df['life_expec'].quantile(0.99)
df['life_expec'][df['life_expec']>= q3_life_expec] = q3_life_expec

q3_income = df['income'].quantile(0.99)
df['income'][df['income']>= q3_income] = q3_income

q3_inflation = df['inflation'].quantile(0.99)
df['inflation'][df['inflation']>= q3_inflation] = q3_inflation

q3_total_fer = df['total_fer'].quantile(0.99)
df['total_fer'][df['total_fer']>= q3_total_fer] = q3_total_fer
```

Clustering

Clustering has been performed using both K-mean and Hierarchical Algorithm with 3 clusters

Scaling : Minmax() scaling has been performed on dataset to bring whole data in same scale

K-means Clustering :

- For K-mean Clustering value of k has been decided based on elbow curve and silhouette score. Based on these 2 approach value of k has been used as 3.

- 3 cluster has been formed based on below criteria

1) Cluster 0: high export, high health, high imports, high income, low inflation, high life_expec, low total_fer, high gdpp, low child_mort

2) Cluster 1: low export, low health, low imports, low income, high inflation, low life_expec, high total_fer, low gdpp, high child_mort

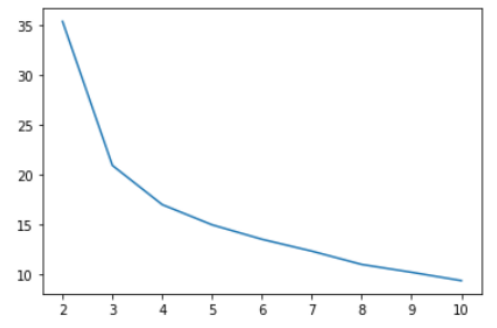
3) Cluster 2: avg export, avg health, avg imports, avg income, avg inflation, avg life_expec, avg total_fer, avg gdpp, avg child_mort

No of countries in each cluster: Below are available count of countries in each cluster

Cluster ID	No of coutries	Variables
0	29	high export, high health, high imports, high income, low inflation, high life_expec, low total_fer, high gdpp, low child_mort
1	46	low export, low health, low imports, low income, high inflation, low life_expec, high total_fer, low gdpp, high child_mort
2	92	avg export, avg health, avg imports, avg income, avg inflation, avg life_expec, avg total_fer, avg gdpp, avg child_mort

Clustering Continues…….

Elbow Curve

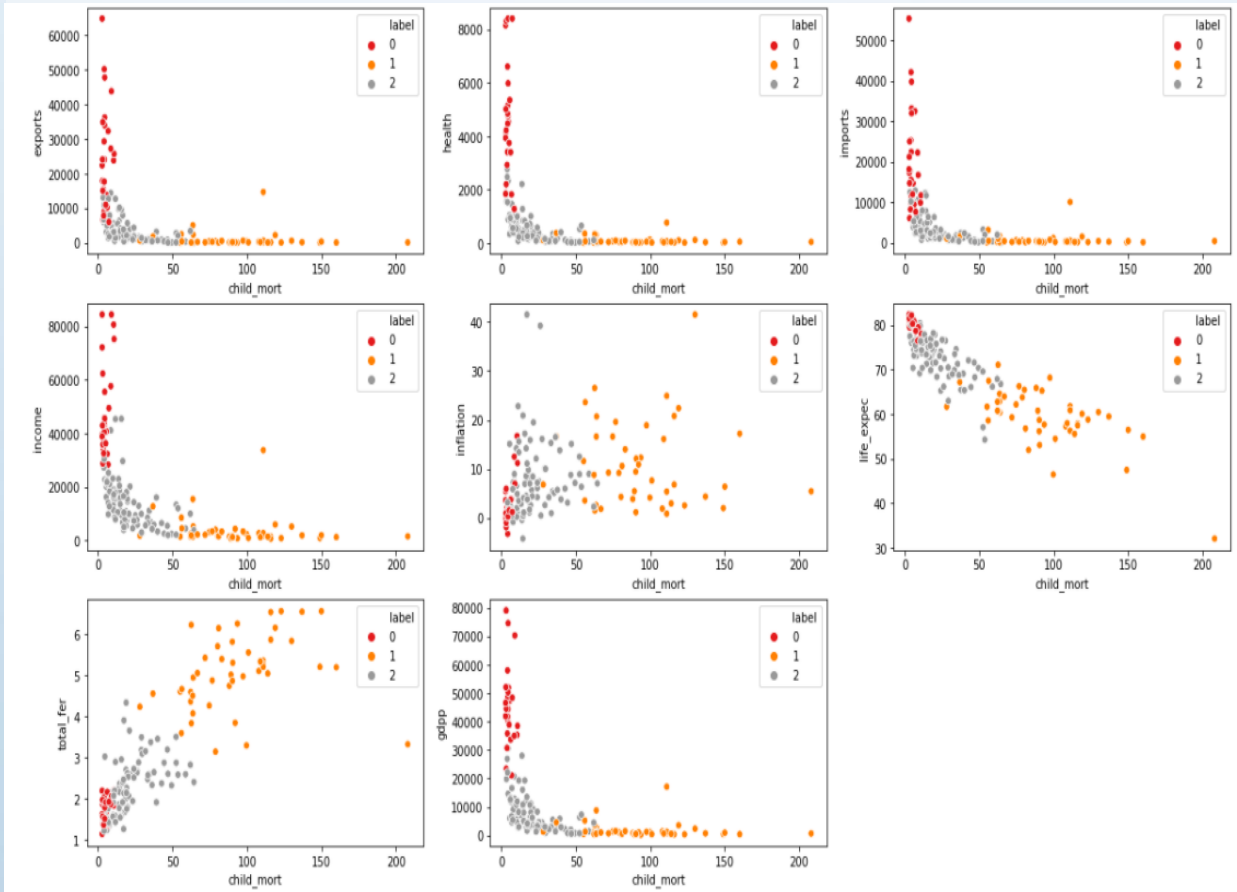


Based on elbow curve optimal value of k is 3

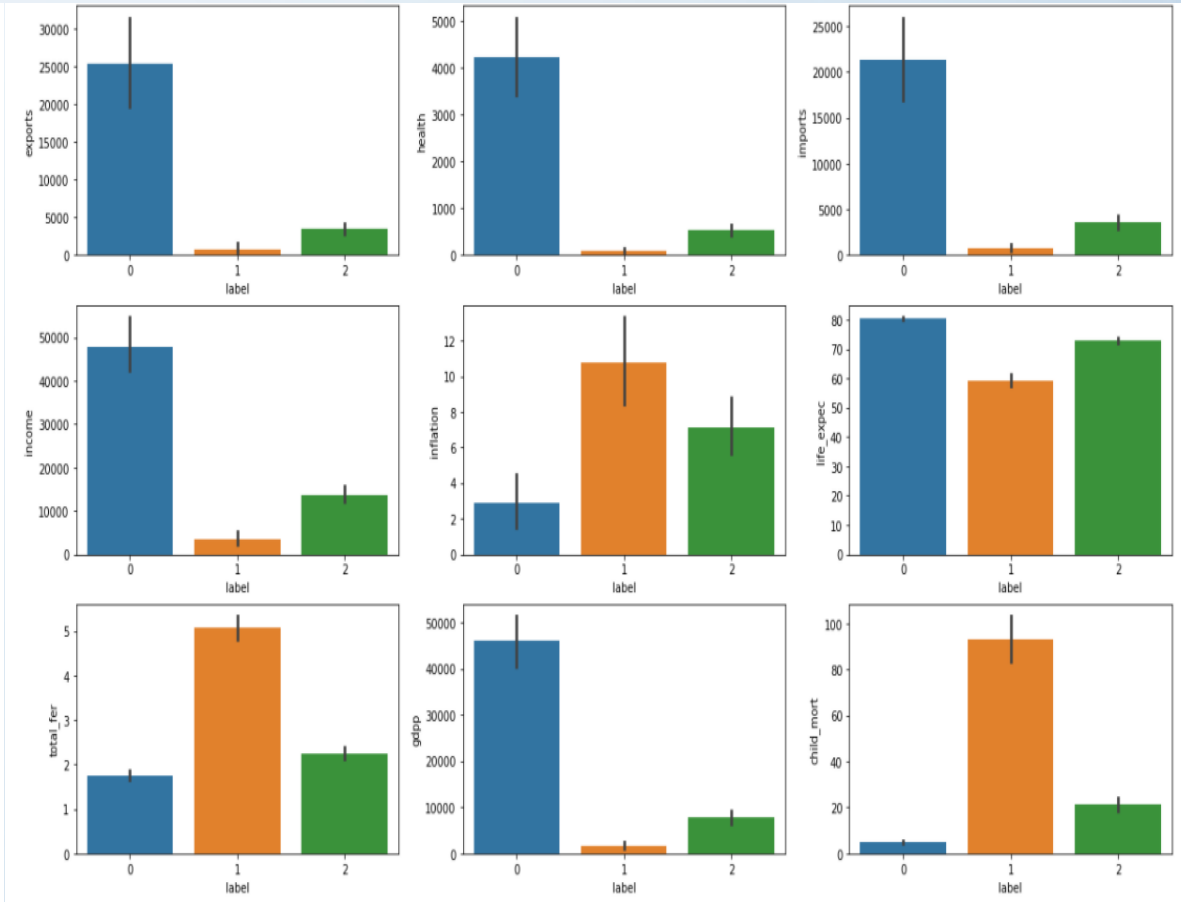
Based on silhouette score and business point of view optimal value of k is 3 as after 3 its decreasing or constant

Silhouette Score

```
For n_clusters=2, the silhouette score is 0.46141999028186625
For n_clusters=3, the silhouette score is 0.4416240153826115
For n_clusters=4, the silhouette score is 0.41872363448215877
For n_clusters=5, the silhouette score is 0.3083555366091141
For n_clusters=6, the silhouette score is 0.3009590932699754
For n_clusters=7, the silhouette score is 0.31501590229263365
For n_clusters=8, the silhouette score is 0.29900889978031636
For n_clusters=9, the silhouette score is 0.2943131745717175
For n_clusters=10, the silhouette score is 0.2649069093872091
```



Formed cluster based on all factors from K-Means



Cluster Visualization with respect to label for K-means

Clustering Continues……..

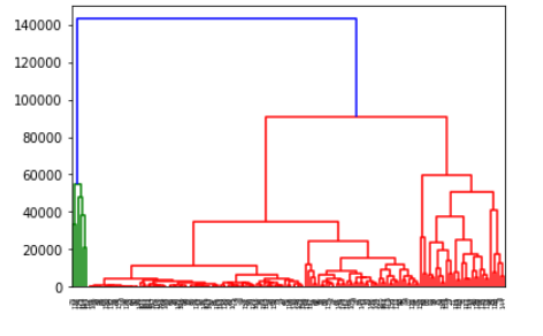
Hierarchical Clustering : 3 clusters have been formed for hierarchical clustering

- For Hierarchical Clustering value of k has been decided based on complete linkage dendrogram. Based on the dendrogram value of k has been used as 3.
- Cluster has been formed based on below criteria

1) Cluster 0: low export, low health, low imports, low income, high inflation, low life_expec, high total_fer, low gdpp, high child_mort

2) Cluster 1: avg export, avg health, avg imports, avg income, avg inflation, avg life_expec, avg total_fer, avg gdpp, avg child_mort

3) Cluster 2: high export, high health, high imports, high income, low inflation, high life_expec, low total_fer, high gdpp, low child_mort

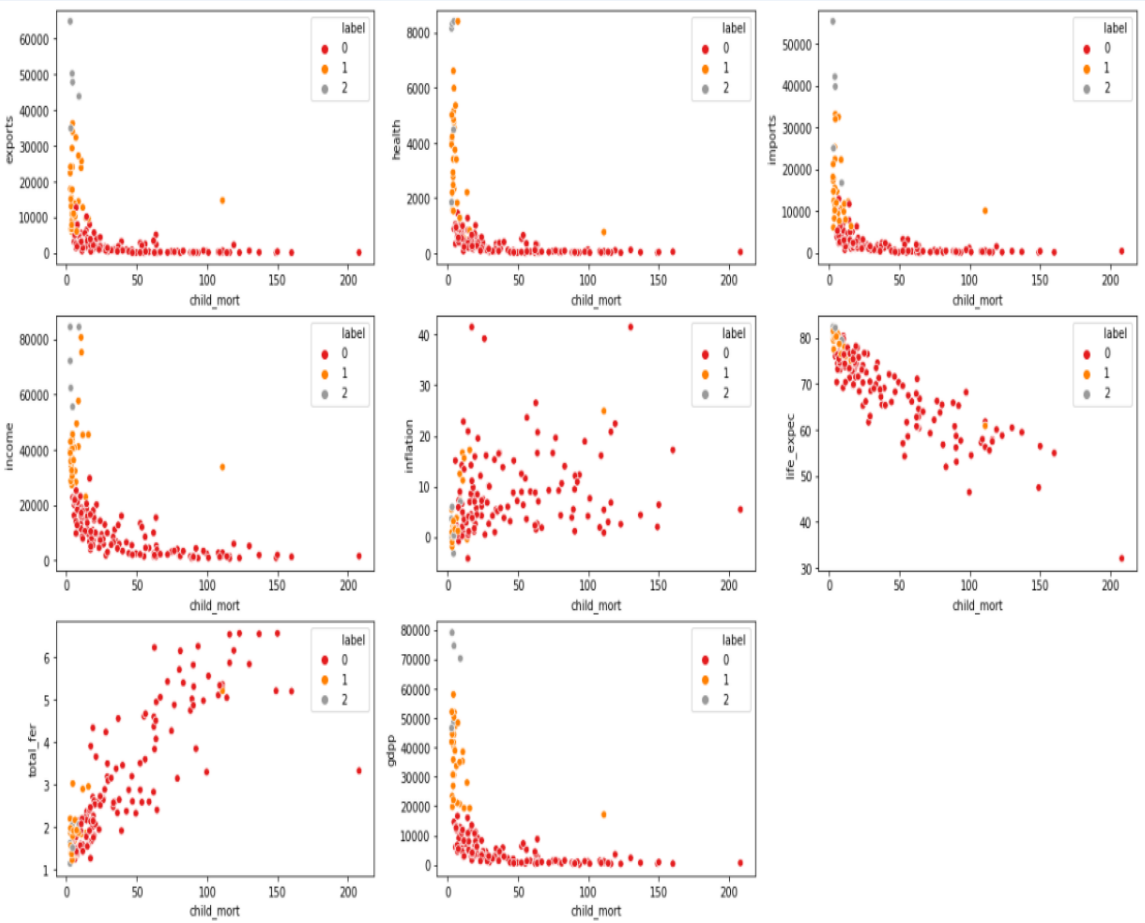


Based on the dendrogram and business point of view optimal value of k is 3

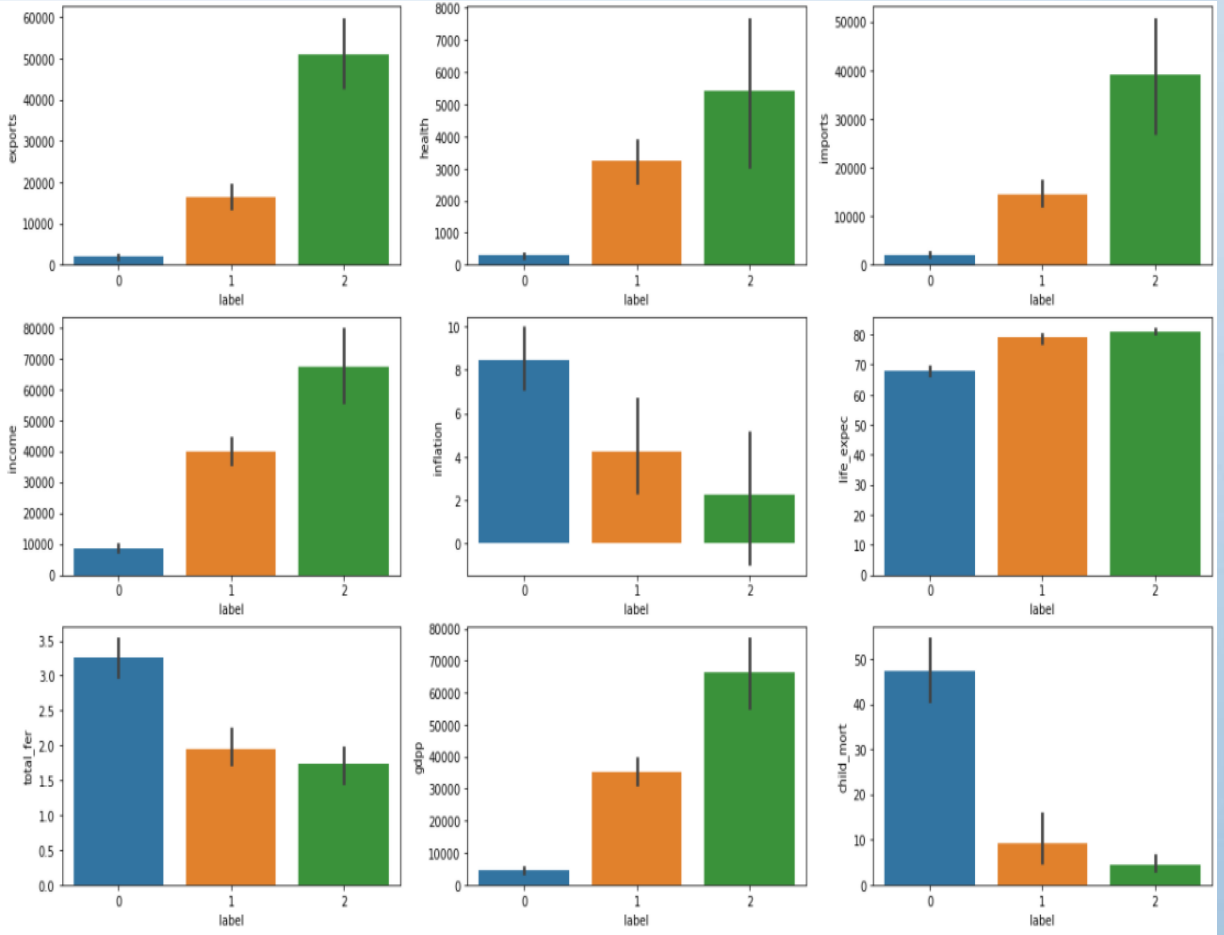
No of countries in each cluster: Below are available count of countries in each cluster

Cluster ID	No of coutries	Variables
0	128	low export, low health, low imports, low income, high inflation, low life_expec, high total_fer, low gdpp, high child_mort
1	33	avg export, avg health, avg imports, avg income, avg inflation, avg life_expec, avg total_fer, avg gdpp, avg child_mort
2	6	high export, high health, high imports, high income, low inflation, high life_expec, low total_fer, high gdpp, low child_mort

Clustering Continues…….



Formed cluster based on all factors from hierarchical



Cluster visualization with respect to labels from hierarchical

Cluster Visualization

Cluster performed via K-means Clustering

1) Cluster 0: high export, high health, high imports, high income, low inflation, high life_expec, low total_fer, high gdpp, low child_mort

2) Cluster 1: low export, low health, low imports, low income, high inflation, low life_expec, high total_fer, low gdpp, high child_mort

3) Cluster 2: avg export, avg health, avg imports, avg income, avg inflation, avg life_expec, avg total_fer, avg gdpp, avg child_mort

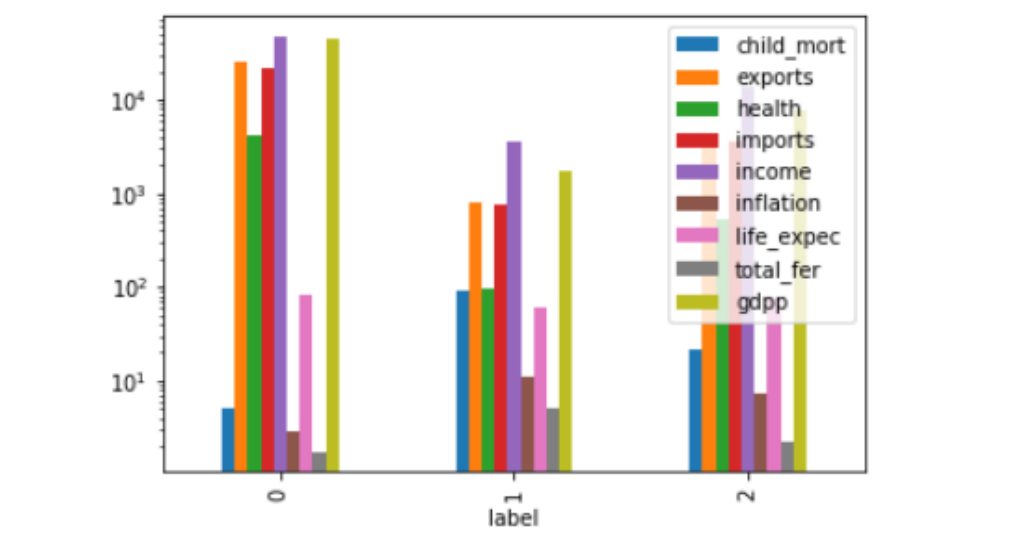
Cluster performed via Hierarchical Clustering

1) Cluster 0: low export, low health, low imports, low income, high inflation, low life_expec, high total_fer, low gdpp, high child_mort

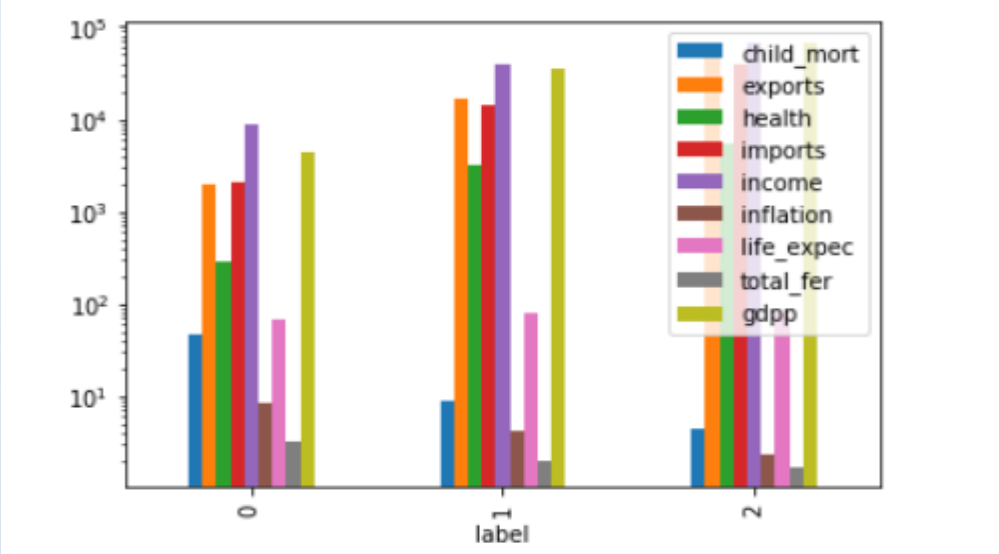
2) Cluster 1: avg export, avg health, avg imports, avg income, avg inflation, avg life_expec, avg total_fer, avg gdpp, avg child_mort

3) Cluster 2: high export, high health, high imports, high income, low inflation, high life_expec, low total_fer, high gdpp, low child_mort

K-Mean Cluster

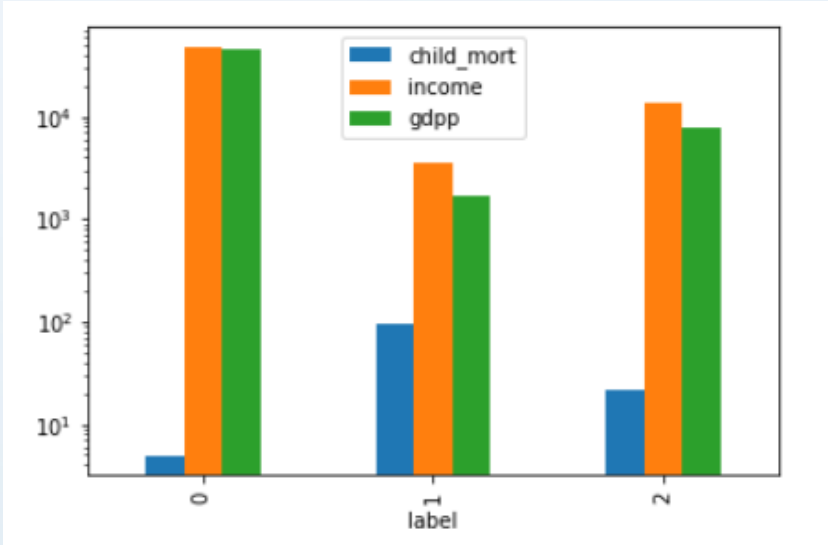
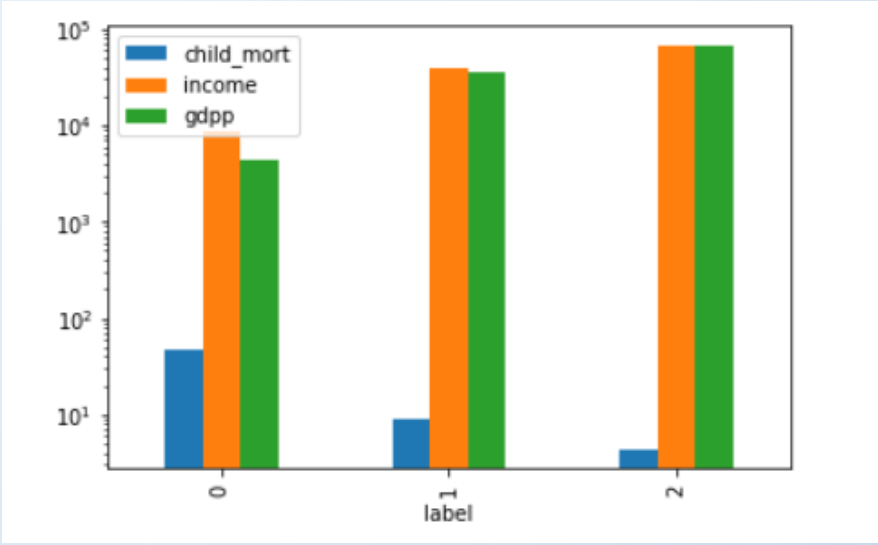


Hierarchical Cluster



Cluster Profiling

Cluster profiling has been done for 3 main factors gdpp, income and health to get the countries, are in dire need of aid

K-means cluster profiling	Hierarchical cluster profiling
	
Countries available in label 1 are in dire need of aid as for those child_mort is very high and gdpp, income is low	Countries available in label 1 are in dire need of aid as for those child_mort is very high and gdpp, income is low

Conclusion

In terms of no countries available in all clusters created via K-Means and Hierarchical Clustering, K-means is having

relative balanced no of records. Hence for final approach will consider K-means algorithm

Proposal

As per the K-means clustering companies are having high child mortality, low income and low gdpp are in dire need of aid.

Hence NGO should invest their raised money to the countries available for label 1 which are having high child mortality rate, low income and low gdpp.

Top 5 companies which are in dire need according to k-Means are as below

Haiti	Sierra Leone	Chad	Central African Republic	Mali

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	label
66	Haiti	208.0	101.286	45.7442	428.314	1500.0	5.45	32.1	3.3300	662.0	1
132	Sierra Leone	160.0	67.032	52.2690	137.655	1220.0	17.20	55.0	5.2000	399.0	1
32	Chad	150.0	330.096	40.6341	390.195	1930.0	6.39	56.5	6.5636	897.0	1
31	Central African Republic	149.0	52.628	17.7508	118.190	888.0	2.01	47.5	5.2100	446.0	1
97	Mali	137.0	161.424	35.2584	248.508	1870.0	4.37	59.5	6.5500	708.0	1

Thank you