Clustering Assignment

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 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.

Business Goal:

- Our job is to categories 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.

Step 1: Reading and understanding the data.

- Reading the Country-data file, and inspect the like identifying shape, info, describe, index, etc. to understand the data.

Step 2: Exploratory Data Analysis(EDA).

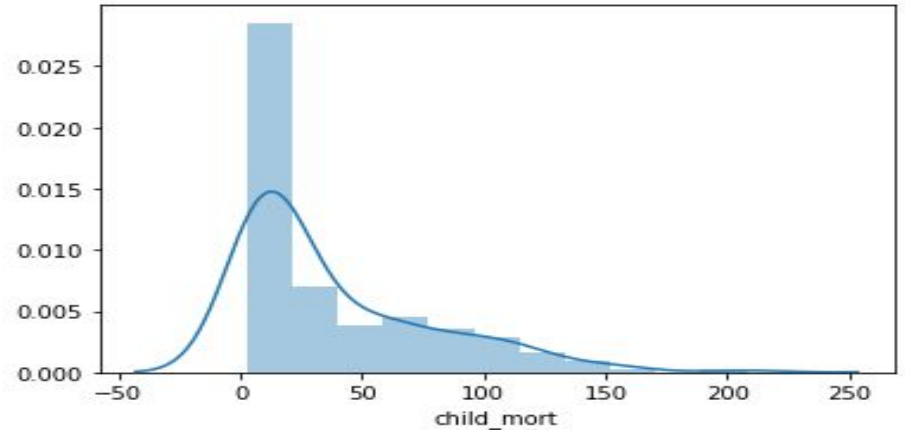
- Clean the data if any missing values.
  - No missing values for the imported data set.
- Standardising the values.
  - We identified that a few variables are given in percentage format. So, converted those variable values into actual values.
- Handling outliers.
  - We observed that there are outliers for few variables and we do have the flexibility of not removing the outliers, so we used capping technique to treat the outliers.

Univariate analysis.

- Visualizing single/individual variables.
- To Understand the each variable frequencies we use distplot and to
- understand the variable we use box plot.

Example for child_mort variable:

```
1  sns.distplot(df['child_mort'])
2  plt.show()
```



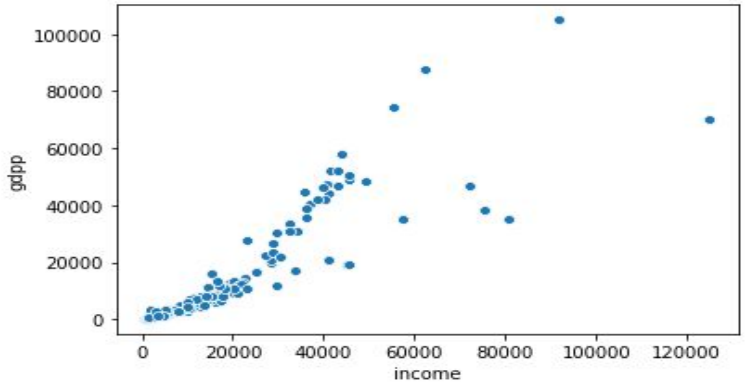
- We observed that the frequency of Death children under 5 years of age per 1000 live births is high at 0 to 50.
- Likewise, visualized for all the variables and understood the variables individually.

Bivariate analysis

- Visualizing multiple variables.
- In bivariate analysis we Understand the each variable relation with other variable.

Example for income Vs. gdpp:

```
1  sns.scatterplot('income','gdpp',data=df)
2  plt.show()
```



- We observed that the variable income increases the gdpp also gradually increasing. So, there is a good correlation between this variables.
- Likewise, we visualized for all the variables understood the variable relations.

Step 3: Hopkins Statistics Test

Hopkins Score Calculation

```
for i in range(10):
    print("Hopikins value is:" , hopkins(df.drop('country', axis=1)))
```

```
Hopikins value is: 0.8881144622725661
Hopikins value is: 0.8936255702945607
Hopikins value is: 0.9403158326144837
Hopikins value is: 0.9028428556305683
Hopikins value is: 0.9115495319452458
Hopikins value is: 0.9178616874097715
Hopikins value is: 0.9386370842626143
Hopikins value is: 0.8309722299225853
Hopikins value is: 0.9077620067834387
Hopikins value is: 0.9073623065842963
```

- Iterations is 88%.
- 88% is best indication for cluster tendency.
- Our data has cluster tendency, So we can go for clustering techniques.

Step 4: Feature Scaling

- We used Standardization scaling here and scaling is for only numeric variables.
- Let’s see the data Before scaling

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
0	Afghanistan	90.2	55.30	41.92	248.30	1610	9.44	56.2	5.82	553
1	Albania	16.6	1145.20	267.90	1987.74	9930	4.49	76.3	1.65	4090
2	Algeria	27.3	1712.64	185.98	1400.44	12900	16.10	76.5	2.89	4460
3	Angola	119.0	2199.19	100.60	1514.37	5900	22.40	60.1	6.16	3530
4	Antigua and Barbuda	10.3	5551.00	735.66	7185.80	19100	1.44	76.8	2.13	12200

After Scaling:

	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
0	1.291532	-0.569622	-0.566956	-0.598740	-0.851668	0.264996	-1.619092	1.902882	-0.702259
1	-0.538949	-0.473858	-0.440391	-0.413584	-0.386946	-0.372073	0.647866	-0.859973	-0.498726
2	-0.272833	-0.424000	-0.486272	-0.476100	-0.221053	1.122143	0.670423	-0.038404	-0.477434
3	2.007808	-0.381249	-0.534091	-0.463973	-0.612045	1.932958	-1.179234	2.128151	-0.530950
4	-0.695634	-0.086742	-0.178410	0.139728	0.125254	-0.764610	0.704258	-0.541946	-0.032042

- Standardization prevents variables with larger scales from dominating how clusters are defined. It allows all variables to be considered by the algorithm with equal importance.

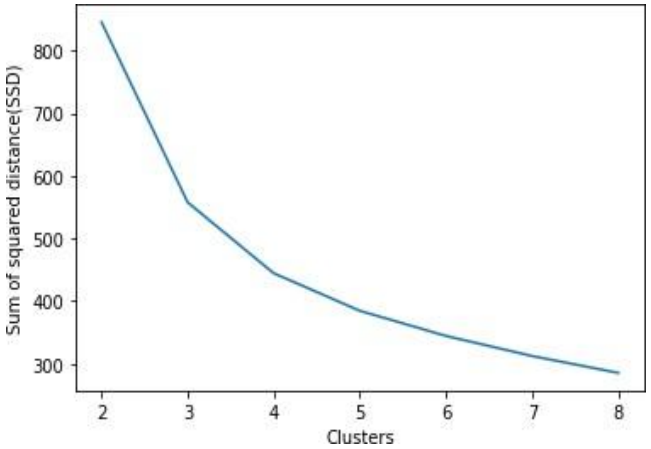
Step 4: Clustering Techniques

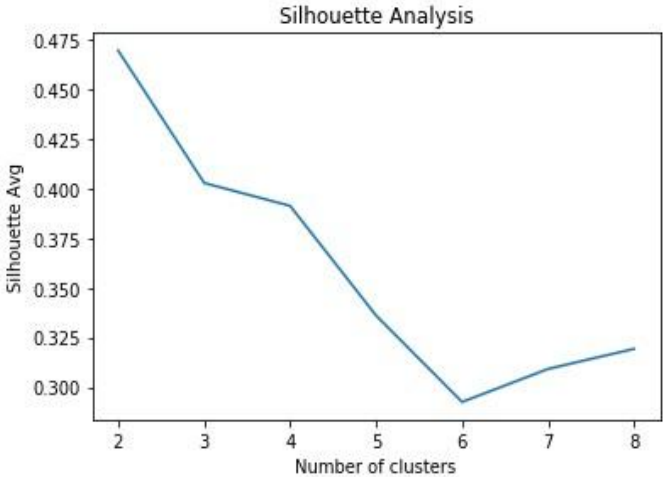
Here are 2 types of Clustering Techniques:

- K-means Clustering:
  - K-means clustering is one of the simplest and popular unsupervised machine learning algorithms, to achieve the optimal K value.
- Hierarchical Clustering:
  - Hierarchical cluster analysis is an algorithm that groups similar objects into groups called clusters.

K-means Clustering:

- In K Means clustering to identify the optimal number of clusters, we have 2 methods.
- Elbow Curve/Sum of Squared Distance(SSD) Silhouette Analysis.

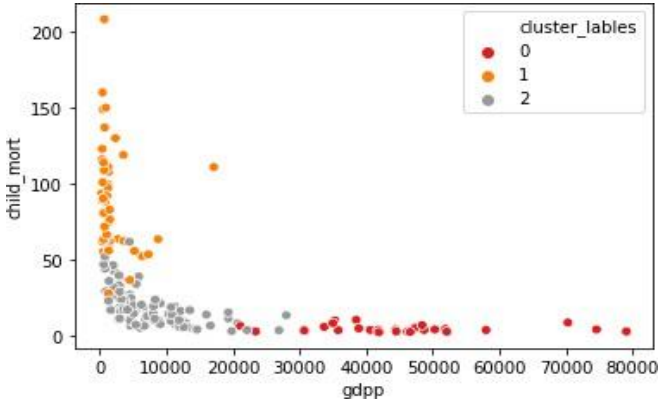


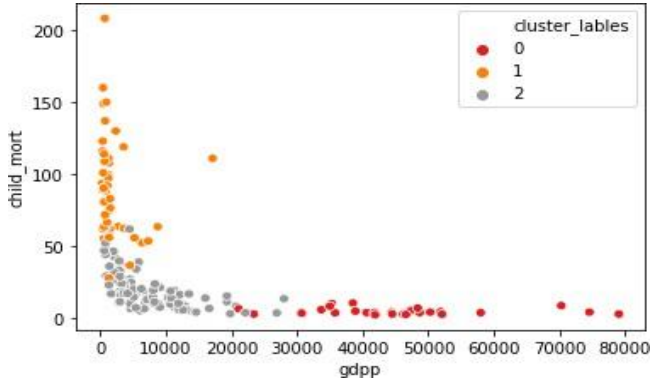


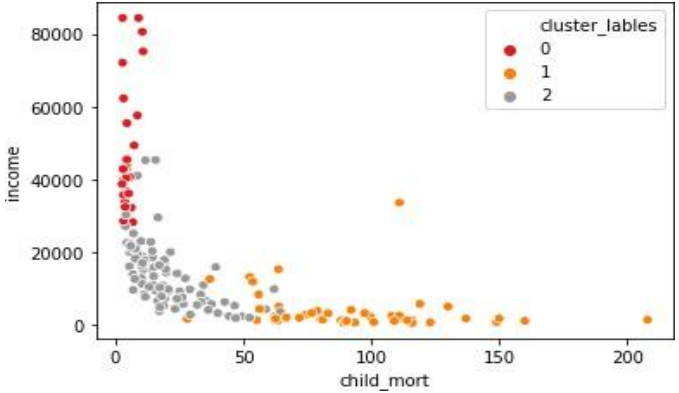
- The value of k at which improvement in distortion declines the most is called the elbow, at which we should stop dividing the data into further clusters.
- So, by observing the elbow curve there is a declines at 3 and 4. And by observing the silhouette scores the average is good at 2 and 3.
- From both the methods we observed and to make the cluster strong we go with 3 clusters and countries are split into 3 clusters.

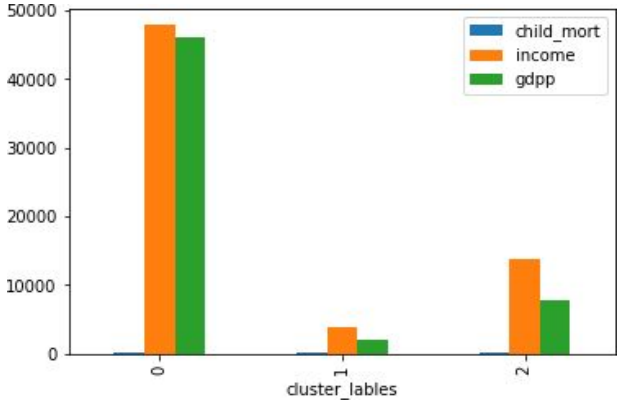
Plot the Clusters

- After countries are split into 3 clusters, plot the variables based on the required/important variables to see the cluster formations.
- And finally in Cluster profiling, by comparing how these three variables - [gdpp, child_mort and income] and we can identify the best clusters.


cluster_lables
0
1
2
child_mort
gdpp
0
50
100
150
200
0
10000
20000
30000
40000
50000
60000
70000
80000


cluster_lables
0
1
2
child_mort
gdpp
0
50
100
150
200
0
10000
20000
30000
40000
50000
60000
70000
80000


cluster_lables
0
1
2
income
child_mort
0
20000
40000
60000
80000
0
50
100
150
200


child_mort
income
gdpp
0
10000
20000
30000
40000
50000
0
1
2
cluster_lables

Final Analysis:

- Final Country list Preparation based on high child_mort, low income and low gdpp are comes under cluster-1 and identify the top 5 countries that ones which are in dire need of aid.

```
1 best_cluster.sort_values(by=['child_mort','income','gdpp'], ascending=[False,True,True]).head()
```

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	cluster_lables
66	Haiti	208.0000	101.2900	45.7400	428.3100	1500	5.4500	32.1000	3.3300	662	1
132	Sierra Leone	160.0000	67.0300	52.2700	137.6600	1220	17.2000	55.0000	5.2000	399	1
32	Chad	150.0000	330.1000	40.6300	390.2000	1930	6.3900	56.5000	6.5900	897	1
31	Central African Republic	149.0000	52.6300	17.7500	118.1900	888	2.0100	47.5000	5.2100	446	1
97	Mali	137.0000	161.4200	35.2600	248.5100	1870	4.3700	59.5000	6.5500	708	1

- The above top 5 countries(Haiti, Sierra Leone, Chad, Central African Republic and Mali) are the one which are in dire need of aid.

Hierarchical Clustering:

To minimise the pairwise distance between the data points, we use the other method called linkage.

There are 2 types of linkages.

- Single Linkage

The shortest distance between points in the clusters called single linkage.

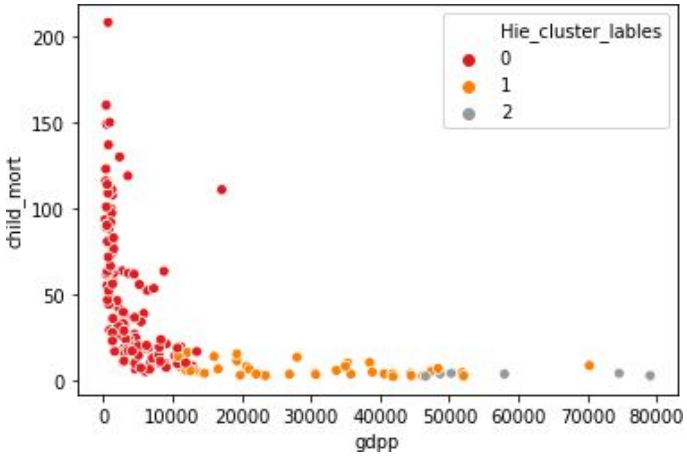
- complete Linkage

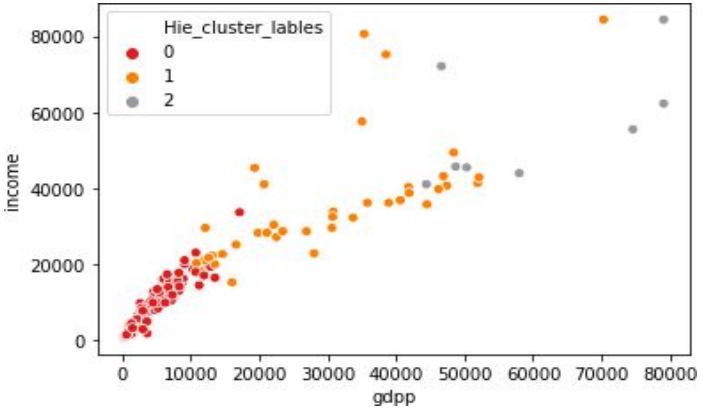
  The maximum distance between any two points in the clusters is called complete linkage.

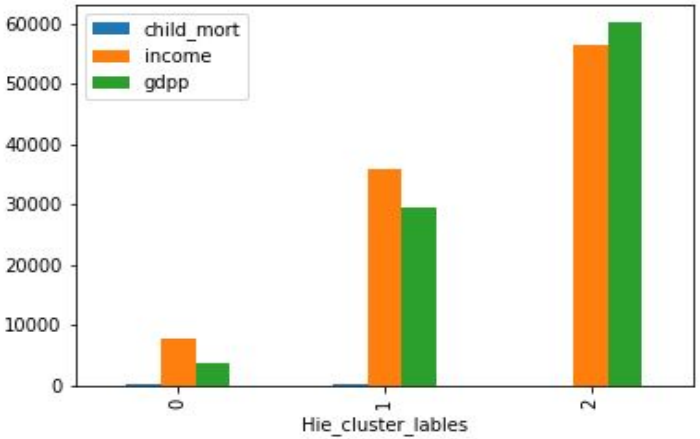
- After observing the dendrograms(A dendrogram is a diagram that shows the hierarchical relationship between objects.) of both single and complete linkages, the best of linkage is complete linkage, because it is the maximum distance between any two points in the clusters.
- Finally by observing the dendrogram we choose clusters as 3 and clusters were formed and countries are split into clusters.
- Both clustering techniques gives the same list of countries. However, Hierarchical clustering is time consuming and needs a lot of processing power because of we compute the distance of each point from each other point.

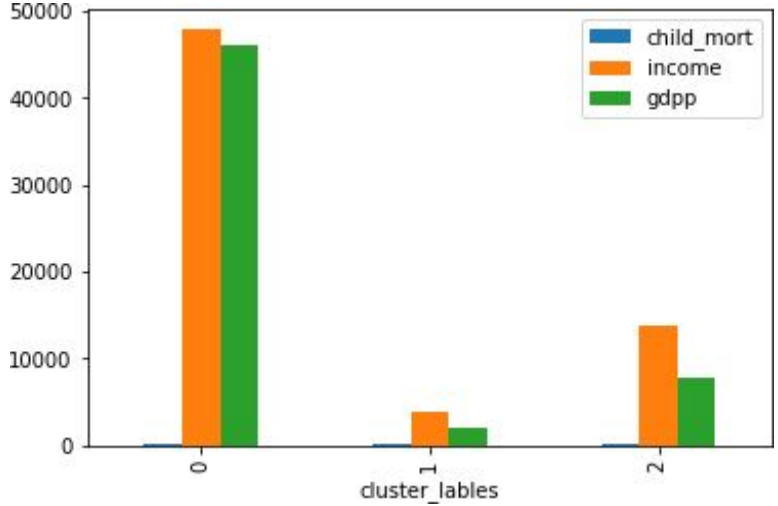
Plot the Clusters:

- After countries are split into 3 clusters, plot the variables based on the required/important variables to see the cluster formations.
- And finally in Cluster profiling, by comparing how these three variables - [gdpp, child_mort and income] and we can identify the best clusters.


Hie_cluster_lables
0
1
2
child_mort
gdpp
0
50
100
150
200
0
10000
20000
30000
40000
50000
60000
70000
80000


Hie_cluster_lables
0
1
2
income
gdpp
0
20000
40000
60000
80000
0
10000
20000
30000
40000
50000
60000
70000
80000


child_mort
income
gdpp
Hie_cluster_lables
0
10000
20000
30000
40000
50000
60000
0
1
2


child_mort
income
gdpp
cluster_lables
0
10000
20000
30000
40000
50000
0
1
2

Final Analysis:

- Final Country list Preparation based on high child_mort, low income and low gdpp are comes under cluster-0 and identify the top 5 countries that ones which are in dire need of aid .

```
1 Hie_best_cluster.sort_values(by=['child_mort','income','gdpp'], ascending=[False,True,True]).head()
```

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	cluster_lables	Hie_cluster_lables
66	Haiti	208.0000	101.2900	45.7400	428.3100	1500	5.4500	32.1000	3.3300	662	1	0
132	Sierra Leone	160.0000	67.0300	52.2700	137.6600	1220	17.2000	55.0000	5.2000	399	1	0
32	Chad	150.0000	330.1000	40.6300	390.2000	1930	6.3900	56.5000	6.5900	897	1	0
31	Central African Republic	149.0000	52.6300	17.7500	118.1900	888	2.0100	47.5000	5.2100	446	1	0
97	Mali	137.0000	161.4200	35.2600	248.5100	1870	4.3700	59.5000	6.5500	708	1	0

- The above top 5 countries(Haiti, Sierra Leone, Chad, Central African Republic and Mali) are the one which are in dire need of aid.

Recommendations and Conclusion :

- The top 5 countries that required help the most are listed below.
  - Haiti
  - Sierra Leone
  - Chad
  - Central African Republic
  - Mali

- The above listed countries are having
  - Low gdpp
    - The GDP per capita. Calculated as the Total GDP divided by the total population.
  - Low income
    - Net income per person.
  - High child_mort
    - Death of children under 5 years of age per 1000 live births.

- So, it is clear and highly recommended that the above listed 5 countries required quick aid.

Thank you