ASSIGNMENT

K-Means Clustering for HELP international

Rama Mishra

OVERVIEW

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 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.

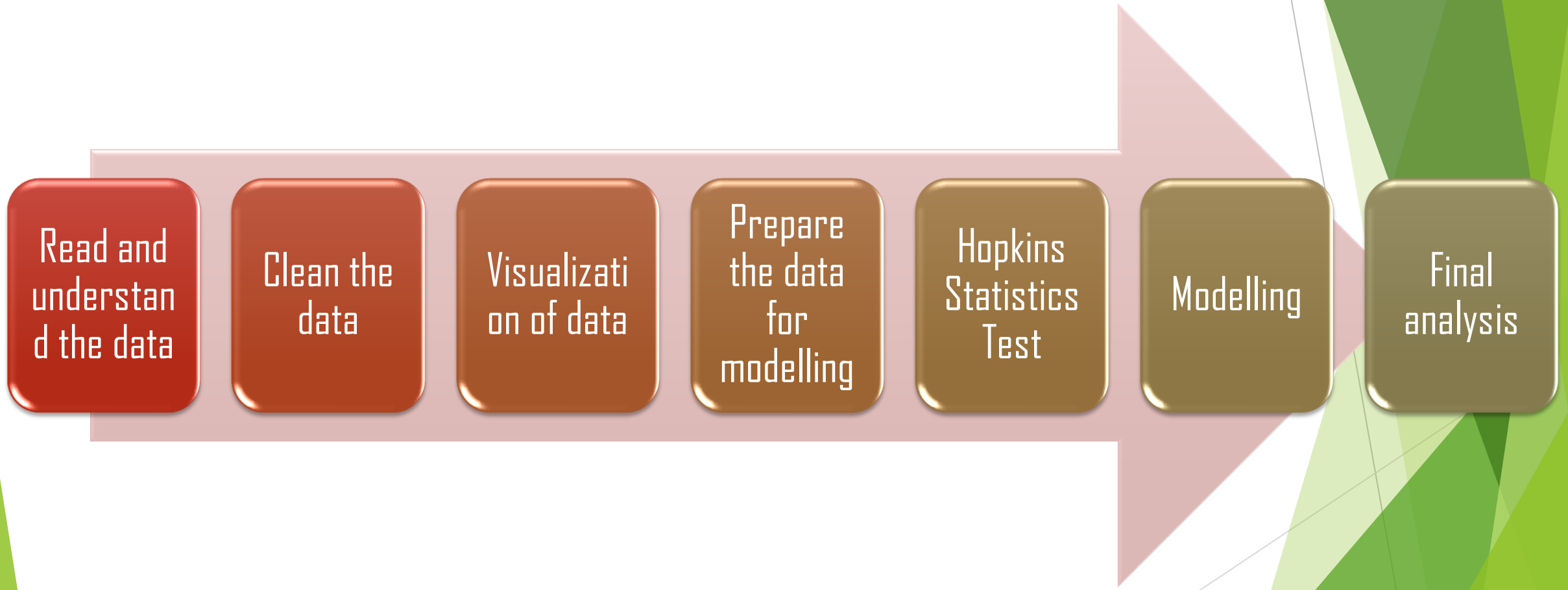
The significant issues that come while making this decision are mostly related to choosing the countries that are in the direst need of aid.

Outcome / Analysis

To categorise the countries using some socio-economic and health factors that determine the overall development of the country.

We need to suggest the countries which the CEO needs to focus on the most.

STEPS/ METHODOLOGY



STEP 1 & 2 DATA UNDERSTANDING & CLEANING

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
0	Afghanistan	90.2	10.0	7.58	44.9	1610	9.44	56.2	5.82	553
1	Albania	16.6	28.0	6.55	48.6	9930	4.49	76.3	1.65	4090
2	Algeria	27.3	38.4	4.17	31.4	12900	16.10	76.5	2.89	4460
3	Angola	119.0	62.3	2.85	42.9	5900	22.40	60.1	6.16	3530
4	Antigua and Barbuda	10.3	45.5	6.03	58.9	19100	1.44	76.8	2.13	12200
...	...	...	...	...	...	...	...	...	...	...
162	Vanuatu	29.2	46.6	5.25	52.7	2950	2.62	63.0	3.50	2970
163	Venezuela	17.1	28.5	4.91	17.6	16500	45.90	75.4	2.47	13500
164	Vietnam	23.3	72.0	6.84	80.2	4490	12.10	73.1	1.95	1310
165	Yemen	56.3	30.0	5.18	34.4	4480	23.60	67.5	4.67	1310
166	Zambia	83.1	37.0	5.89	30.9	3280	14.00	52.0	5.40	1460

167 rows × 10 columns

This fig shows the data has 10 columns and 167 rows.

Data set does not have any missing values nor any inconsistent data type.

There is no duplicate values provided in dataset.

Converting the exports, imports and health variables to actual values

Converting imports,exports and health spending from percentage values to actual values of their GDP per capita. Because the percentage values don't give a clear picture of that country. For example Austria and Belarus have almost same exports % (Austria=51.3, Belarus= 51.4) but their gdpp has a huge gap (Austria=46900, Belarus= 6030) which doesn't give an accurate idea of which country is more developed than the other.

```
# Converting exports,imports and health spending percentages to absolute values.
country['exports'] = country['exports'] * country['gdpp']/100
country['imports'] = country['imports'] * country['gdpp']/100
country['health'] = country['health'] * country['gdpp']/100
country
```

	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
...	...	...	...	...	...	...	...	...	...	...
162	Vanuatu	29.2	1384.02	155.9250	1565.190	2950	2.62	63.0	3.50	2970
163	Venezuela	17.1	3847.50	662.8500	2376.000	16500	45.90	75.4	2.47	13500
164	Vietnam	23.3	943.20	89.6040	1050.620	4490	12.10	73.1	1.95	1310
165	Yemen	56.3	393.00	67.8580	450.640	4480	23.60	67.5	4.67	1310
166	Zambia	83.1	540.20	85.9940	451.140	3280	14.00	52.0	5.40	1460

FINDING CORRELATION BETWEEN DIFFERENT VARIABLES

STEP 3 - DATA VISUALISATION

INFERENCE

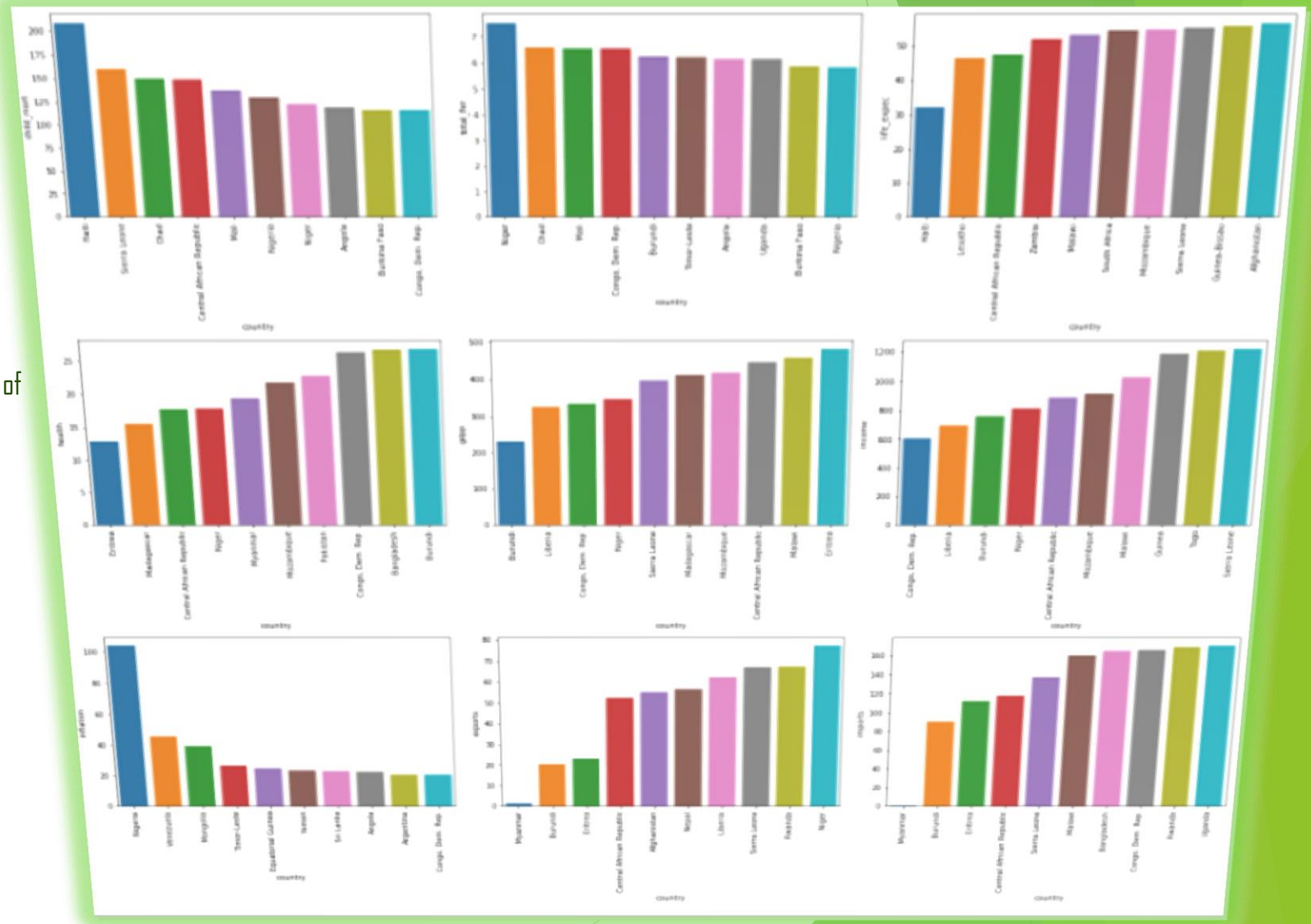
- Imports & Exports are highly correlated with correlation of 0.99
- Health & GDPP are highly correlated with correlation of 0.92
- Income & GDPP are highly correlated with correlation of 0.9
- Child_Mortality & Life Expectancy are highly correlated with correlation of -0.89
- Child_Mortality & Total Fertility are highly correlated with correlation of 0.85
- GDPP & exports are highly correlated with correlation of 0.77
- GDPP & Imports are highly correlated with correlation of 0.76
- Life_Expectency & Total_Fertility are highly correlated with correlation of -0.76
- Income & Exports are highly correlated with correlation of 0.73

UNIVARIATE ANALYSIS OF THE DATA

We need to choose the countries that are in the direct need of aid.

Hence, we need to identify those countries with using some socio-economic and health factors that determine the overall development of the country.

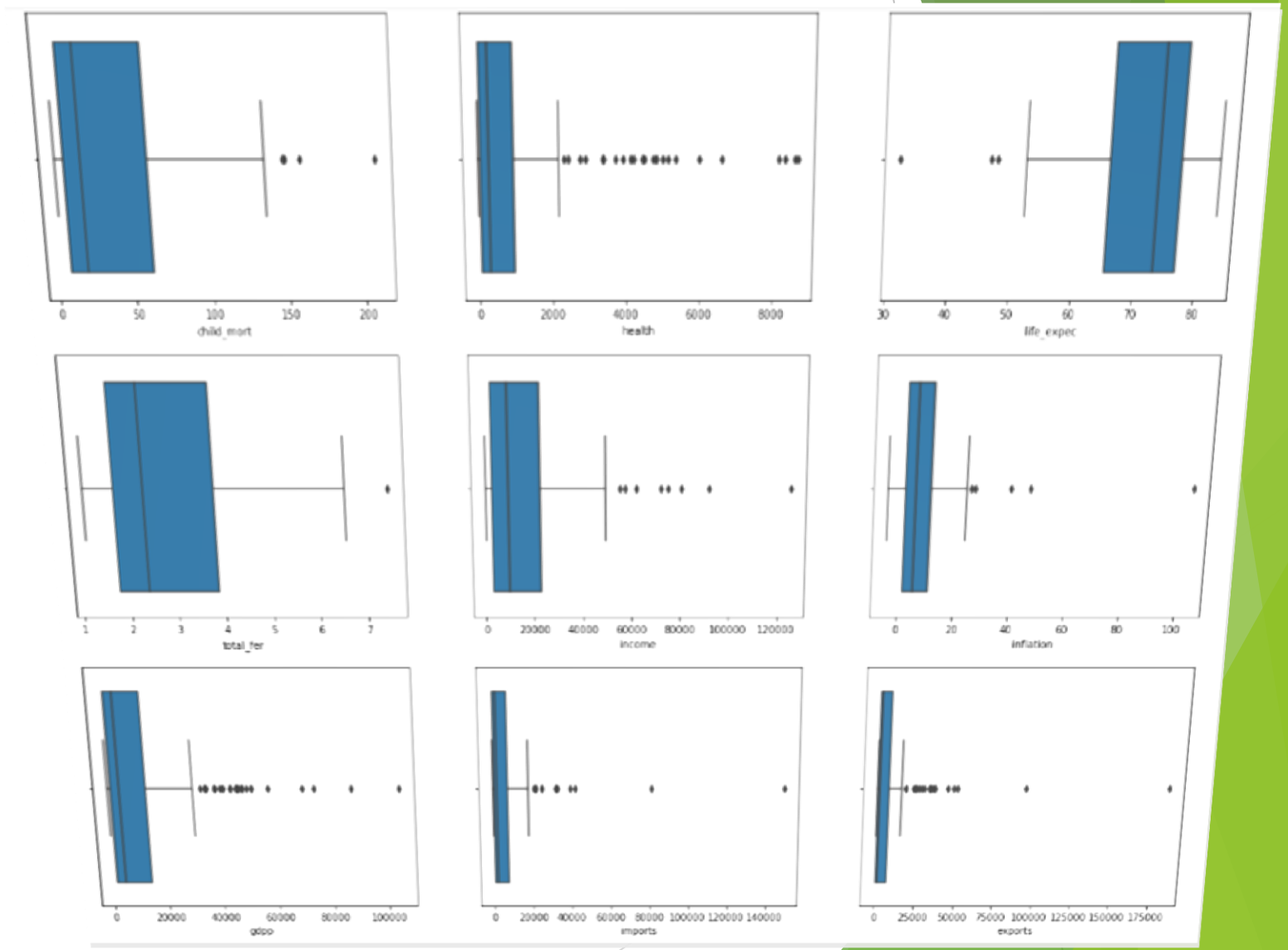
Top 10 countries based on every column



UNIVARIATE ANALYSIS OF THE DATA

BOX PLOT

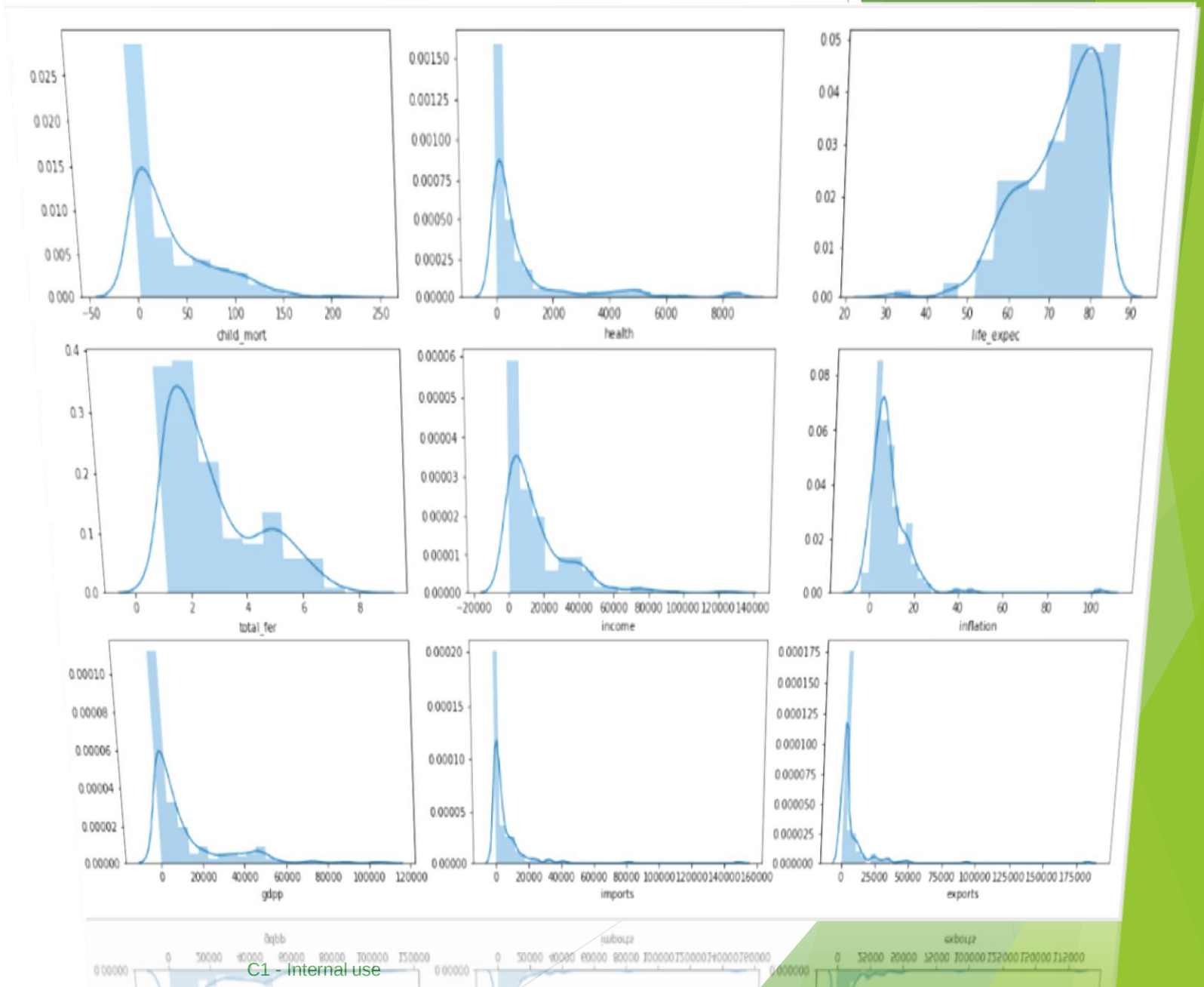
We can see some outliers in all variables



UNIVARIATE ANALYSIS OF THE DATA

DIST PLOT

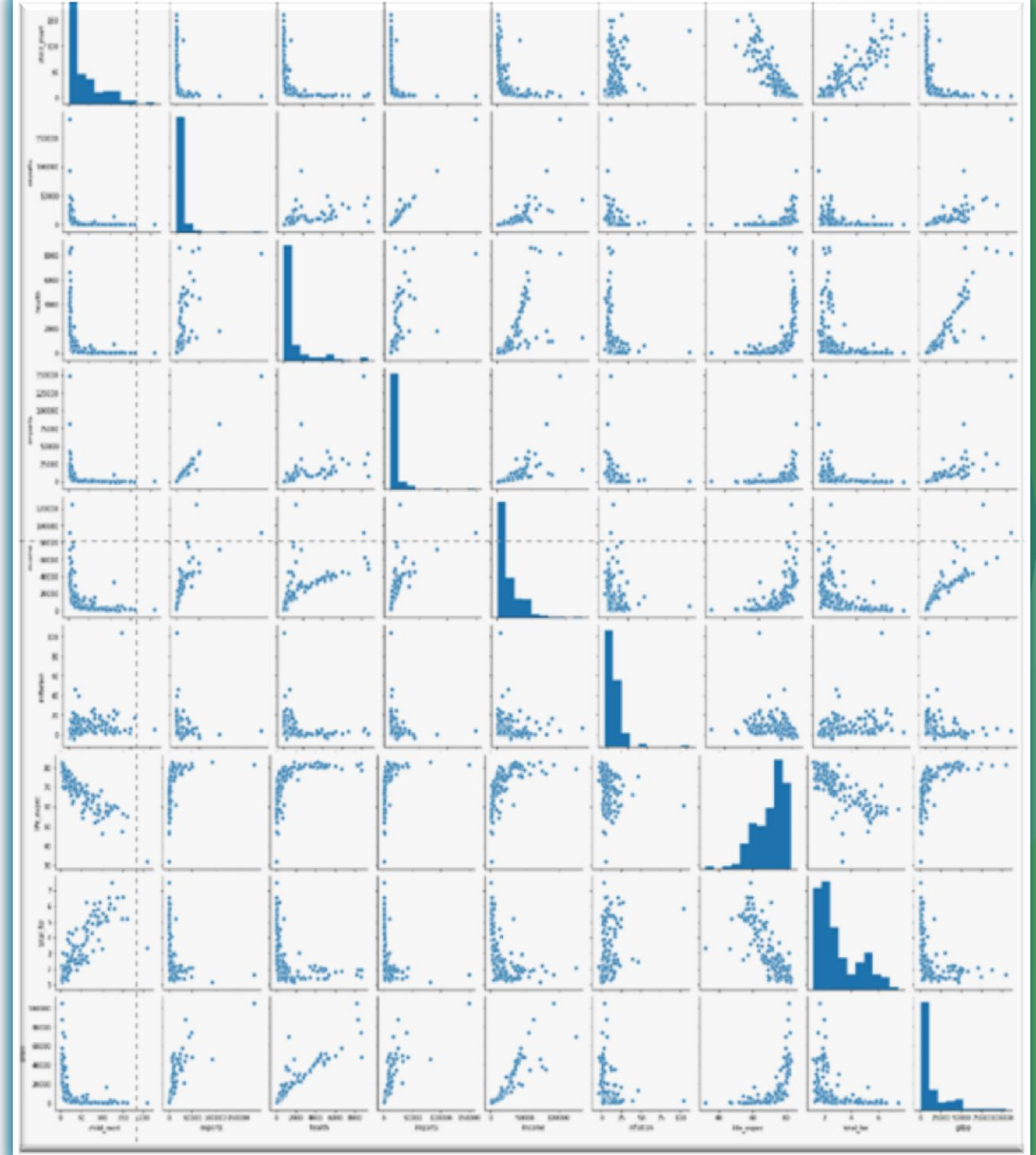
We can see distribution of data



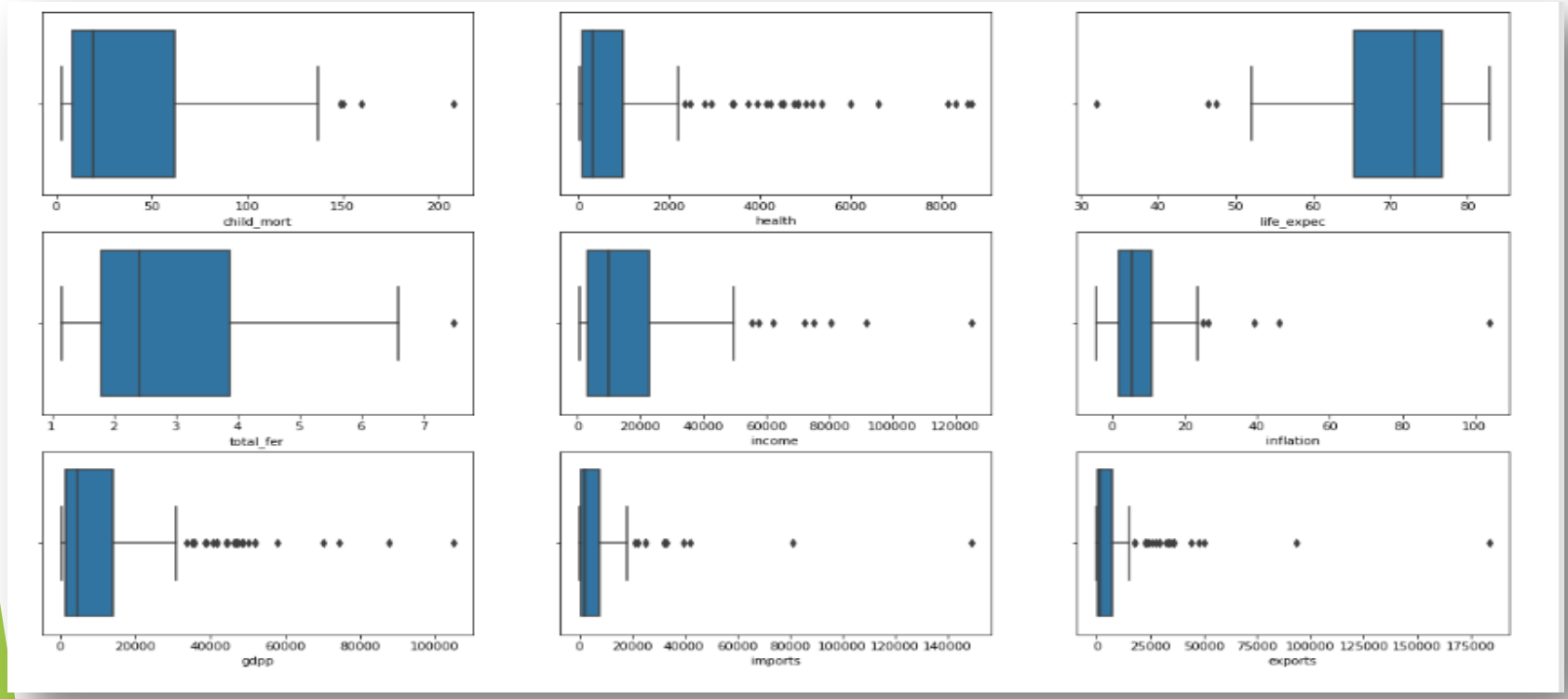
BIVARIATE ANALYSIS OF THE DATA

PAIR PLOT

We can see scatter plots between all the numerical variables



STEP 4:- PREPARING THE DATA FOR MODELLING CHECKING FOR OUTLIERS IN THE DATA AND THEN SCALING IT



I have treated outliers by using capping technique for all the variables.

Then I have scaled the data by Standardizing it.

```
# capping (statistical) the outliers

# outlier treatment for child_mort
Q1 = country1.child_mort.quantile(0.01)
Q4 = country1.child_mort.quantile(0.99)
country1['child_mort'][country1['child_mort'] <= Q1]=Q1
country1['child_mort'][country1['child_mort'] >= Q4]=Q4

# outlier treatment for exports
Q1 = country1.exports.quantile(0.01)
Q4 = country1.exports.quantile(0.99)
country1['exports'][country1['exports'] <= Q1]=Q1
country1['exports'][country1['exports'] >= Q4]=Q4

# outlier treatment for health
Q1 = country1.health.quantile(0.01)
Q4 = country1.health.quantile(0.99)
country1['health'][country1['health'] <= Q1]=Q1
country1['health'][country1['health'] >= Q4]=Q4

# outlier treatment for imports
Q1 = country1.imports.quantile(0.01)
Q4 = country1.imports.quantile(0.99)
country1['imports'][country1['imports'] <= Q1]=Q1
country1['imports'][country1['imports'] >= Q4]=Q4

# outlier treatment for income
Q1 = country1.income.quantile(0.01)
Q4 = country1.income.quantile(0.99)
country1['income'][country1['income'] <= Q1]=Q1
country1['income'][country1['income'] >= Q4]=Q4
# outlier treatment for life_expec
Q1 = country1.life_expec.quantile(0.01)
Q4 = country1.life_expec.quantile(0.99)
country1['life_expec'][country1['life_expec'] <= Q1]=Q1
country1['life_expec'][country1['life_expec'] >= Q4]=Q4

# outlier treatment for total_fer
Q1 = country1.total_fer.quantile(0.01)
Q4 = country1.total_fer.quantile(0.99)
country1['total_fer'][country1['total_fer'] <= Q1]=Q1
country1['total_fer'][country1['total_fer'] >= Q4]=Q4

# outlier treatment for gdpp
Q1 = country1.gdpp.quantile(0.01)
Q4 = country1.gdpp.quantile(0.99)
country1['gdpp'][country1['gdpp'] <= Q1]=Q1
country1['gdpp'][country1['gdpp'] >= Q4]=Q4
```

STEP 5:- HOPKINS STATISTICS TEST

I have performed hopkins test on my scaled data and I got 0. 87 score, which indicates my data is good for clustering

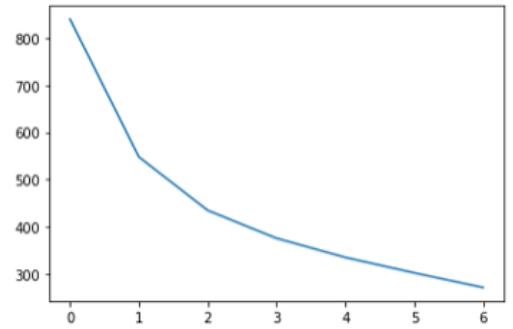
```
# Hopkins score

hopkins(country_scaled)
```

```
0.8727122846411209
```

STEP 6:- MODELLING

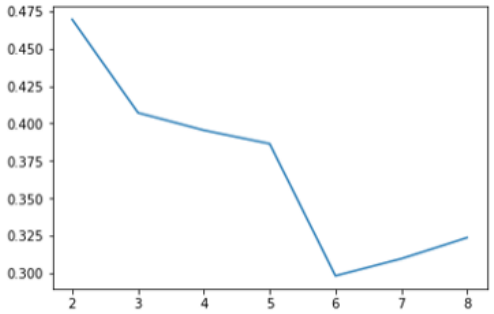
I have selected k=3, based on both the silhouette score and elbow graph, so the number of clusters will be formed in k-mean clustering will be 3.

Elbow graph




```
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 0.38611500797464143
For n_clusters=6, the silhouette score is 0.28636340993161424
For n_clusters=7, the silhouette score is 0.2916485283562354
For n_clusters=8, the silhouette score is 0.2855078623647604
```

Silhouette score graph



CLUSTER PROFILING

From the business understanding we have learnt that

Child_Mortality, Income, Gdpp are some important factors which decides the development of any country. Hence, we will proceed with cluster profiling by using these 3 variables.

GDPP: (The GDP per capita) Calculated as the Total GDP divided by the total population.

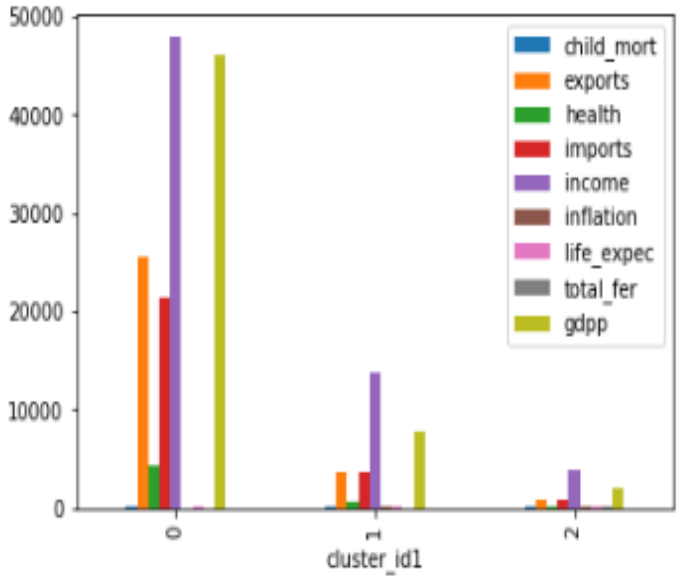
Child_mort: Death of children under 5 years of age per 1000 live births.

Income: Net income per person.

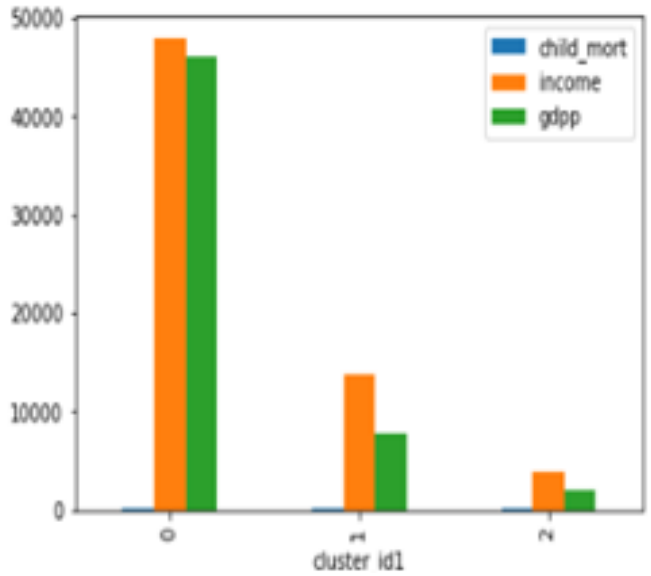
K-MEAN METHOD (CLUSTER PROFILING)

- By using k=3, Analysis of clusters formed from k mean on basis of 3 columns i.e income, GDPP and Child_mort

```
country1.groupby('cluster_id1').mean().plot(kind= 'bar')
plt.show()
```




```
country1[['child_mort', 'income', 'gdpp', 'cluster_id1']].groupby('cluster_id1').mean().plot(kind= 'bar')
plt.show()
```

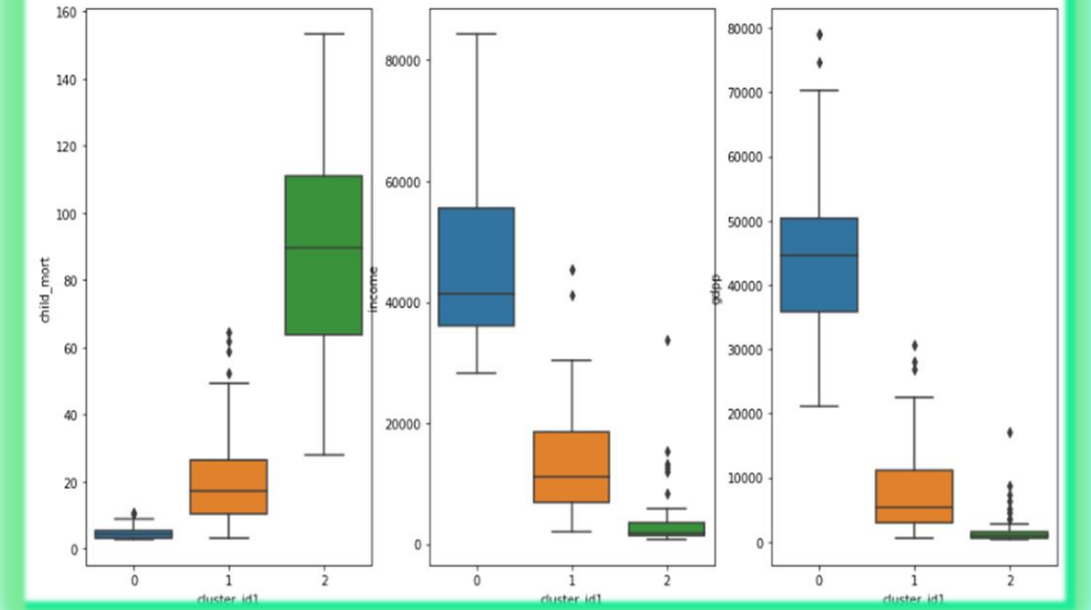


Inference:- Income and Gdpp are measures of development. Higher the per capita income and gdpp better is the country's development. Income per capita and gdpp seems lowest for countries in **clusters 2** and Child Mortality is also highest for **Cluster 2** Hence, these countries are in need of some help.

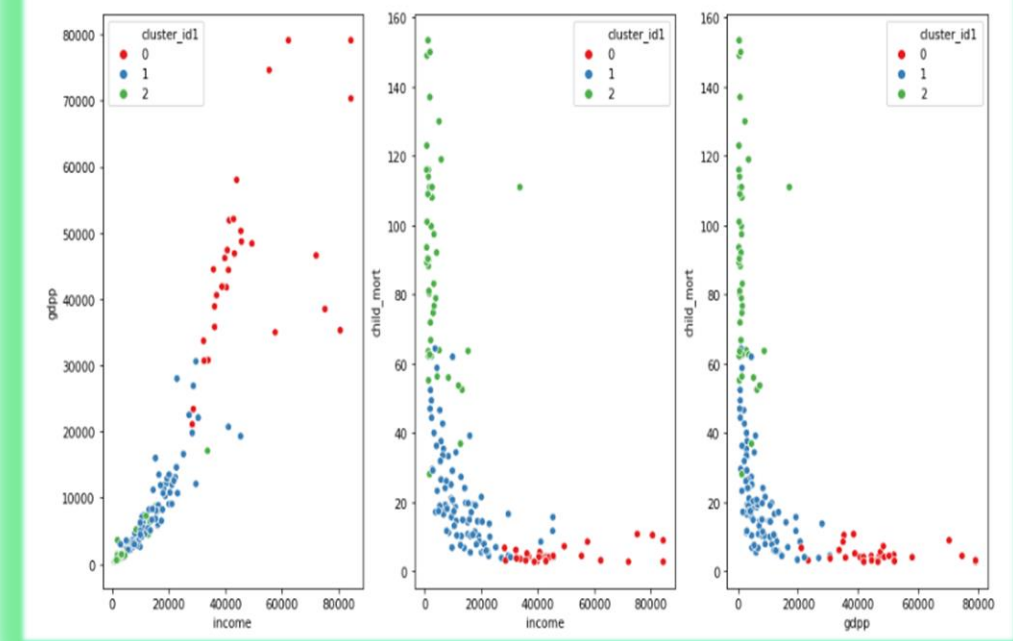
K-MEAN METHOD (CLUSTER PROFILING)

- By using k=3, Analysis of clusters formed from k mean on basis of 3 columns i.e income, GDPP and Child mort

```
plt.figure(figsize=(15,8))
plt.subplot(1,3,1)
sns.boxplot(x='cluster_id1', y='child_mort', data=country1)
plt.subplot(1,3,2)
sns.boxplot(x='cluster_id1', y='income', data=country1)
plt.subplot(1,3,3)
sns.boxplot(x='cluster_id1', y='gdpp', data=country1)
plt.show()
```




```
sns.scatterplot(x= 'income', y= 'gdpp', hue='cluster_id1', legend='full', data=country1, palette= 'Set1')
plt.subplot(1,3,2)
sns.scatterplot(x= 'income', y= 'child_mort', hue='cluster_id1', legend='full', data=country1, palette= 'Set1')
plt.subplot(1,3,3)
sns.scatterplot(x= 'gdpp', y= 'child_mort', hue='cluster_id1', legend='full', data=country1, palette= 'Set1')
plt.show()
```



Inference:- Income and Gdpp are measures of development. Higher the per capita income and gdpp better is the country's development. Income per capita and gdpp seems lowest for countries in clusters 2 and Child Mortality is also highest for Cluster 2 Hence, these countries are in need of some help.

K-MEAN METHOD (CLUSTER PROFILING)

- From cluster_id1 cluster2 shows countries with low income, gdpp and high child_mort
- Top 10 countries which are in direct need of aid
- These countries are sorted based on ascending form of income and gdpp and descending form of child_morality

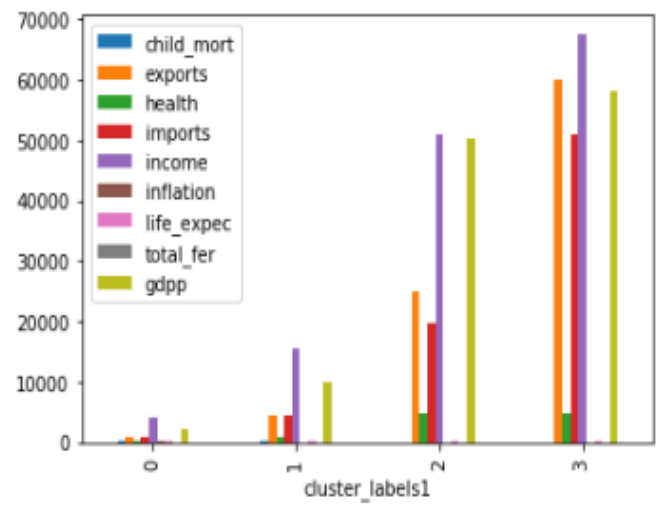
```
select_count=kmean_selected_clust.sort_values(by=['income', 'gdpp', 'child_mort'], ascending=[True, True, False]).head(10)
select_count
```

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	cluster_id1
88	Liberia	89.3	62.457000	38.5860	302.80200	742.24	5.47	60.8	5.0200	331.62	2
37	Congo, Dem. Rep.	116.0	137.274000	26.4194	165.66400	742.24	20.80	57.5	6.5400	334.00	2
26	Burundi	93.6	22.243716	26.7960	104.90964	764.00	12.30	57.7	6.2600	331.62	2
112	Niger	123.0	77.256000	17.9568	170.86800	814.00	2.55	58.8	6.5636	348.00	2
31	Central African Republic	149.0	52.628000	17.7508	118.19000	888.00	2.01	47.5	5.2100	446.00	2
106	Mozambique	101.0	131.985000	21.8299	193.57800	918.00	7.64	54.5	5.5600	419.00	2
94	Malawi	90.5	104.652000	30.2481	160.19100	1030.00	12.10	53.1	5.3100	459.00	2
63	Guinea	109.0	196.344000	31.9464	279.93600	1190.00	16.10	58.0	5.3400	648.00	2
150	Togo	90.3	196.176000	37.3320	279.62400	1210.00	1.18	58.7	4.8700	488.00	2
132	Sierra Leone	153.4	67.032000	52.2690	137.65500	1220.00	17.20	55.0	5.2000	399.00	2

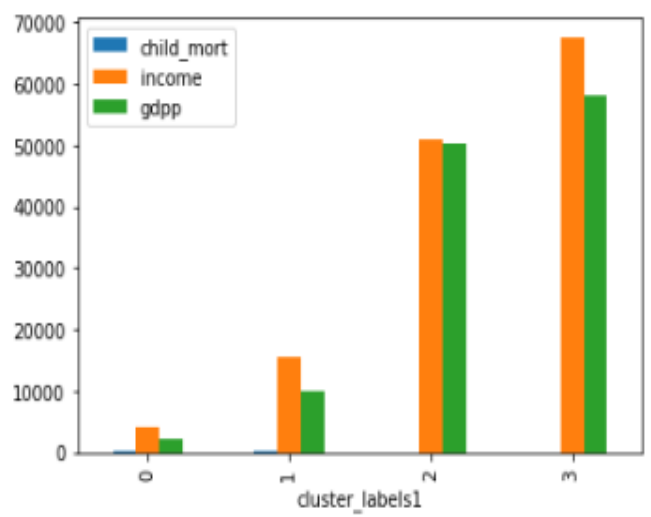
HIERARCHICAL METHOD(COMPLETE LINKAGE) (CLUSTER PROFILING)

By using k=4, Analysis of clusters formed from hierarchical method on basis of 3 columns i.e income, gdpp and child_mort

```
country1.groupby('cluster_labels1').mean().plot(kind= 'bar')
plt.show()
```




```
country1[['child_mort', 'income', 'gdpp', 'cluster_labels1']].groupby('cluster_labels1').mean().plot(kind= 'bar')
plt.show()
```

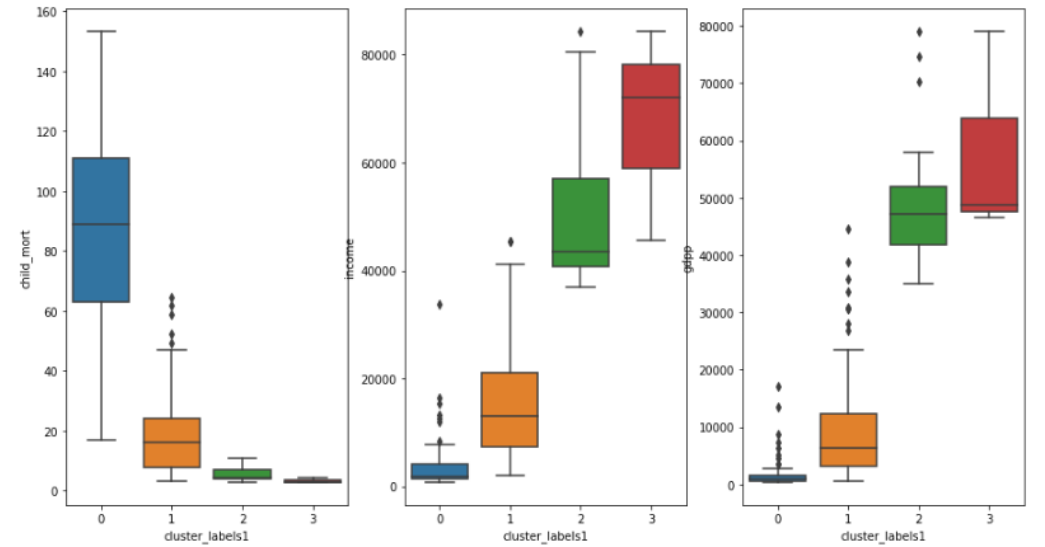


Inference:- Income and Gdpp are measures of development. Higher the per capita income and gdpp better is the country's development. Income per capita and gdpp seems lowest for countries in **clusters 0** and Child Mortality is also highest for **Cluster 0** Hence, these countries are in need of some help.

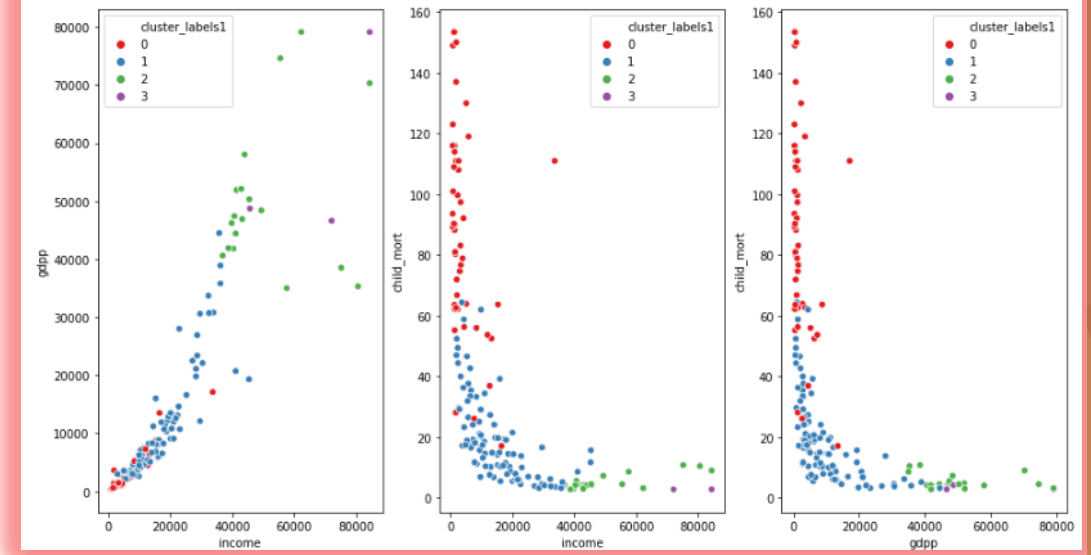
HIERARCHICAL METHOD(COMPLETE LINKAGE) (CLUSTER PROFILING)

By using k=4, Analysis of clusters formed from hierarchical method on basis of 3 columns i.e income, gdpp and child_mort

```
sns.boxplot(x='cluster_labels1', y='child_mort', data=country1)
plt.subplot(1,3,2)
sns.boxplot(x='cluster_labels1', y='income', data=country1)
plt.subplot(1,3,3)
sns.boxplot(x='cluster_labels1', y='gdpp', data=country1)
plt.show()
```




```
sns.scatterplot(x= 'income', y= 'gdpp', hue='cluster_labels1', legend='full', data=country1, palette= 'Set1')
plt.subplot(1,3,2)
sns.scatterplot(x= 'income', y= 'child_mort', hue='cluster_labels1', legend='full', data=country1, palette= 'Set1')
plt.subplot(1,3,3)
sns.scatterplot(x= 'gdpp', y= 'child_mort', hue='cluster_labels1', legend='full', data=country1, palette= 'Set1')
plt.show()
```



Inference:- Income and Gdpp are measures of development. Higher the per capita income and gdpp better is the country's development. Income per capita and gdpp seems lowest for countries in **clusters 0** and Child Mortality is also highest for **Cluster 0** Hence, these countries are in need of some help.

HIERARCHICAL METHOD(COMPLETE LINKAGE) (CLUSTER PROFILING)

From cluster_labels1 cluster0 shows countries with low income, gdpp and high child_mort.

Top 10 countries which are in direct need of aid

These countries are sorted based on ascending form of income and gdpp and descending form of child_morality

```
hier_cont= hier_clust.sort_values(by=['income', 'gdpp', 'child_mort'], ascending=[True, True, False]).head(10)
hier_cont
```

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	cluster_labels1
88	Liberia	89.3	62.457000	38.5860	302.80200	742.24	5.47	60.8	5.0200	331.62	0
37	Congo, Dem. Rep.	116.0	137.274000	26.4194	165.66400	742.24	20.80	57.5	6.5400	334.00	0
26	Burundi	93.6	22.243716	26.7960	104.90964	764.00	12.30	57.7	6.2600	331.62	0
112	Niger	123.0	77.256000	17.9568	170.86800	814.00	2.55	58.8	6.5636	348.00	0
31	Central African Republic	149.0	52.628000	17.7508	118.19000	888.00	2.01	47.5	5.2100	446.00	0
106	Mozambique	101.0	131.985000	21.8299	193.57800	918.00	7.64	54.5	5.5600	419.00	0
94	Malawi	90.5	104.652000	30.2481	160.19100	1030.00	12.10	53.1	5.3100	459.00	0
63	Guinea	109.0	196.344000	31.9464	279.93600	1190.00	16.10	58.0	5.3400	648.00	0
150	Togo	90.3	196.176000	37.3320	279.62400	1210.00	1.18	58.7	4.8700	488.00	0
132	Sierra Leone	153.4	67.032000	52.2690	137.65500	1220.00	17.20	55.0	5.2000	399.00	0

STEP 7:- FINAL ANALYSIS TEST

I have selected **K-mean cluster method** for selection of countries(as the distribution of countries in each cluster was proper and also in k-mean the selection of k for the cluster formation depends on elbow curve method and silhouette score) and based on the information provided by the final clusters I will deduce the final list of
Top 10 countries which are in need of aid.

1. Liberia
2. Congo, Dem. Rep.
3. Burundi
4. Niger
5. Central African Republic
6. Mozambique
7. Malawi
8. Guinea
9. Togo
10. Sierra Leone

```
plt.subplot(1,3,1)
sns.barplot(x='country', y='income', data=select_count)
plt.xticks(rotation = 90)

plt.subplot(1,3,2)
sns.barplot(x='country', y='gdpp', data=select_count)
plt.xticks(rotation = 90)

plt.subplot(1,3,3)
sns.barplot(x='country', y='child_mort', data=select_count)
plt.xticks(rotation = 90)

fig.tight_layout()
plt.show()
```