HELP INTERNATIONAL CLUSTERING ASSIGNMENT

VIJAY TEJA V

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. 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.

DATA HANDLING

- ✓ Percentages have been converted into actual values
- ✓ And as per checks there are no null values

```
# convert exports, health & imports columns in actual amounts from percentage provided
orig_df['exports'] = orig_df['exports'] * orig_df['gdpp'] /100
orig_df['health'] = orig_df['health'] * orig_df['gdpp'] /100
orig_df['imports'] = orig_df['imports'] * orig_df['gdpp'] /100
```

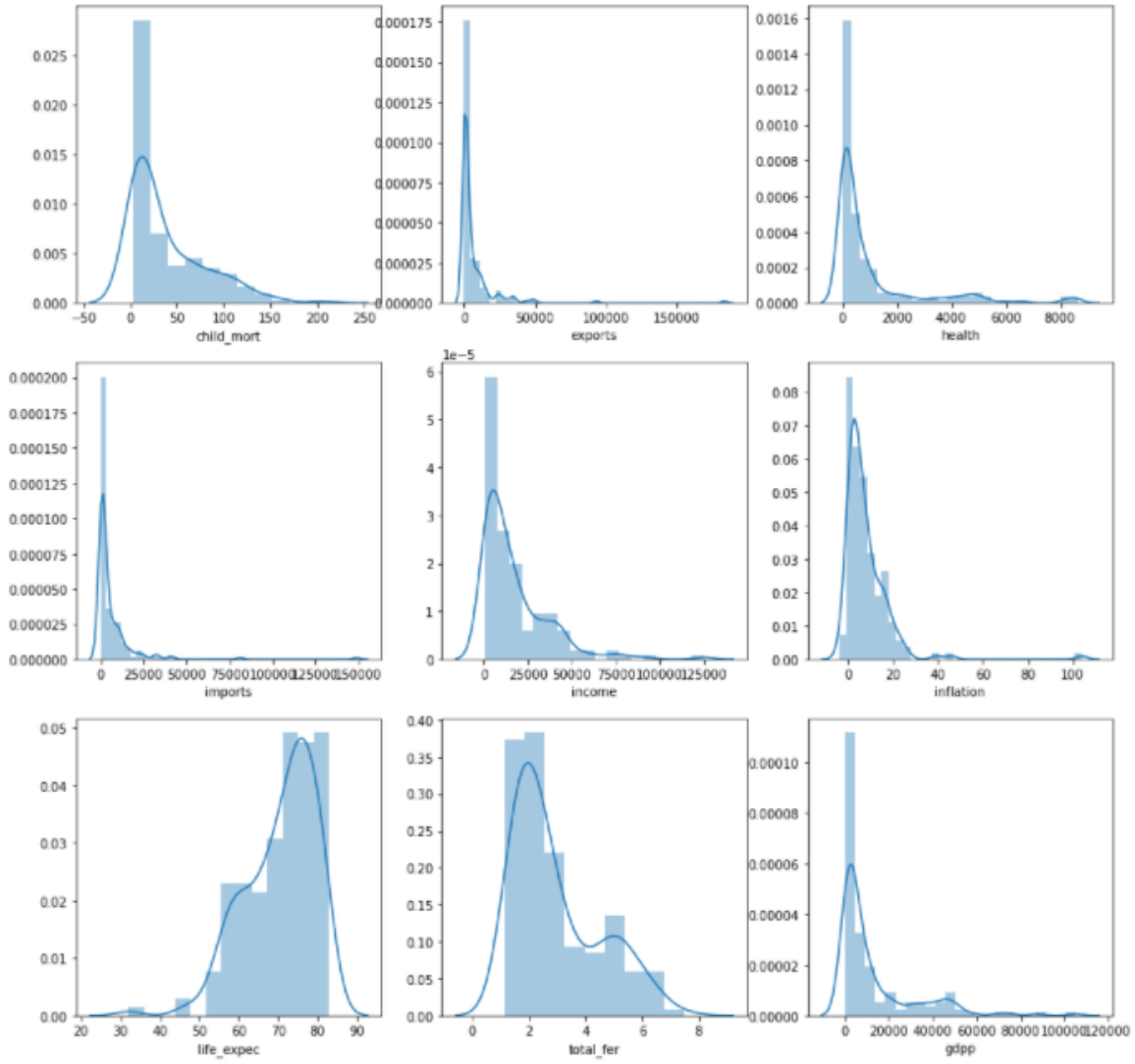
```
#check null
orig_df.isnull().sum()
```

```
country       0
child_mort    0
exports       0
health        0
imports       0
income        0
inflation     0
life_expec    0
total_fer     0
gdpp          0
dtype: int64
```

Exploratory Data Analysis

Univariate Analysis – Distplot

As per the above screenshot, the distribution of data is almost skewed right/left i.e., most of the data is concenrated towards lower economical background (underdeveloped) and some part of data towards high economic background(developed).



Exploratory Data Analysis

BIVARIATE ANALYSIS - Heat Map and Correlation

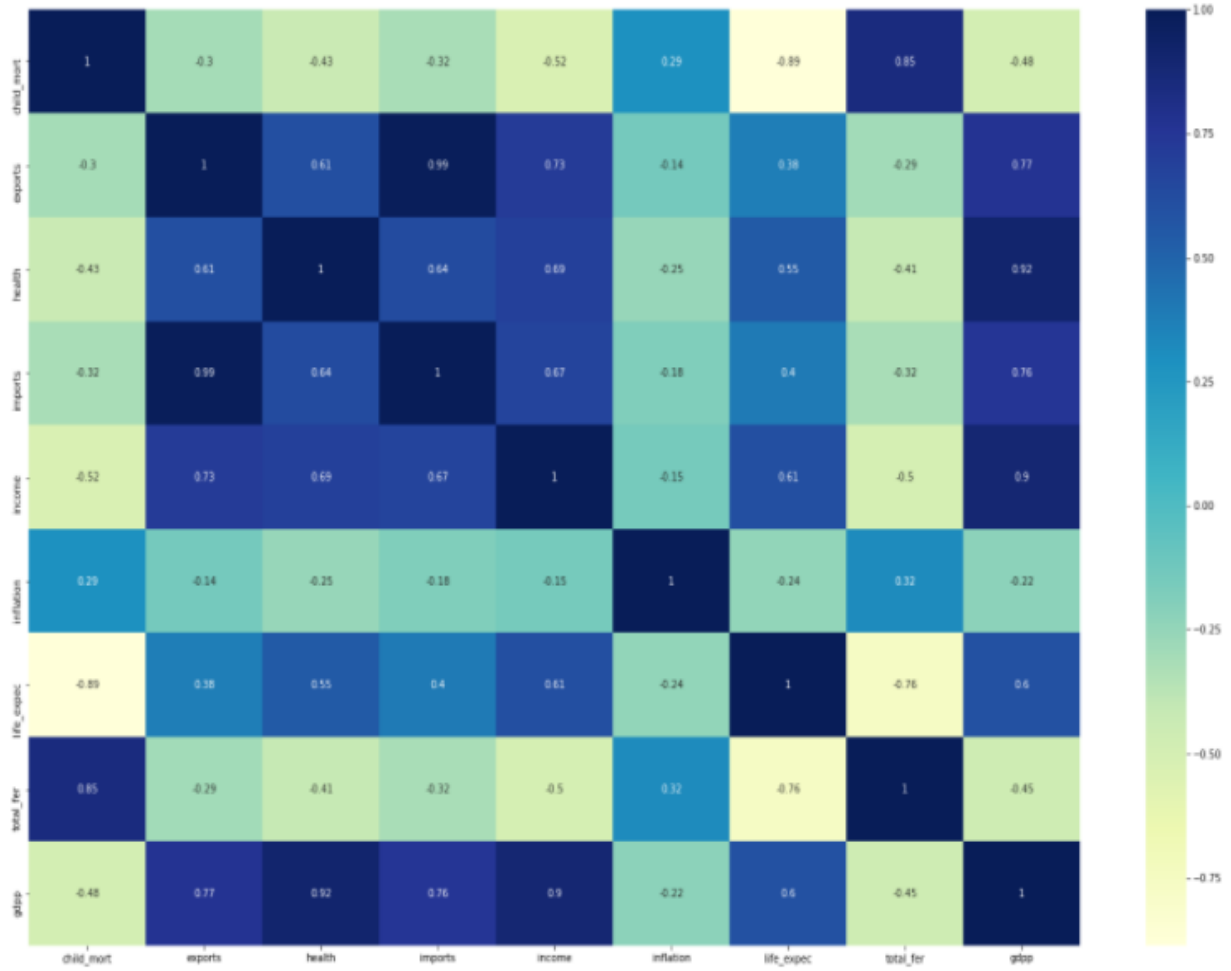
* Better the exports, health, imports & income better the GDP. And we had calculated the actuals of imports, exports and health from gdpp percentages.

* Exports and imports are directly proportional to each other.

* Better the health and income, better the life expectancies.

* Higher the child mortality, higher the total fertility.

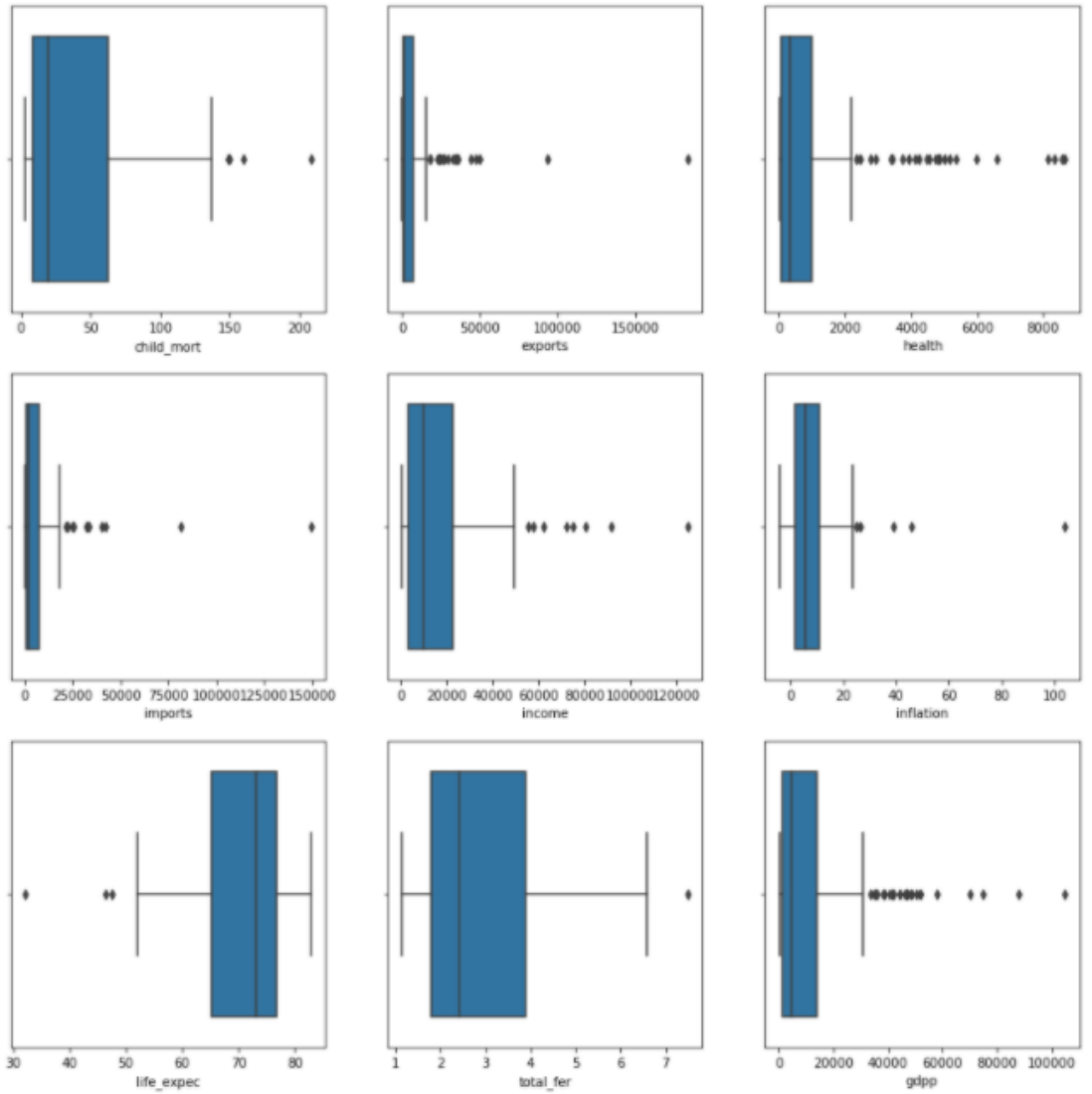
* Higher the child mortaliity, lesser the life expectancies.



Outlier Handling

Child Mortality and Inflation field doesn't require treatment of high end outliers and it doesn't have lower end outliers.

Life Expectancy has low end outlier which could be ignored in outlier handling. Others are treated which includes exports, health, imports, income, total-fer, gdpp by soft or hard capping.



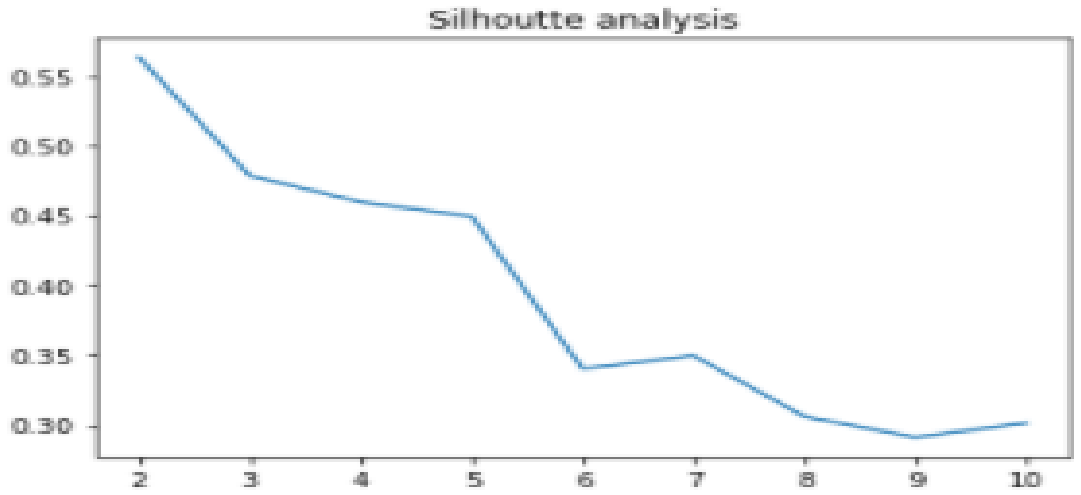
HOPKINS SCORE AND SCALING

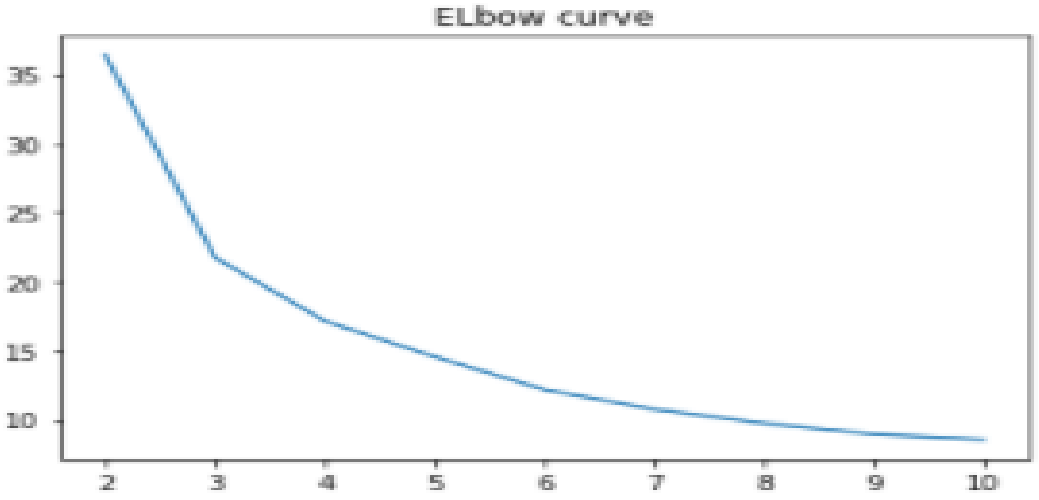
- By running the Hopkins score on the dataset multiple times, we see the values in the range of 86-94%
- Scaling is done by Normalization technique

K- Means clustering

FINDING OPTIMAL K

As per Sumit's suggestion, the Highest value has been taken from Silhoutte graph and the elbow point should be taken for the optimal K value. There by from the graphs, we could conclude the optimal **K is 3**.

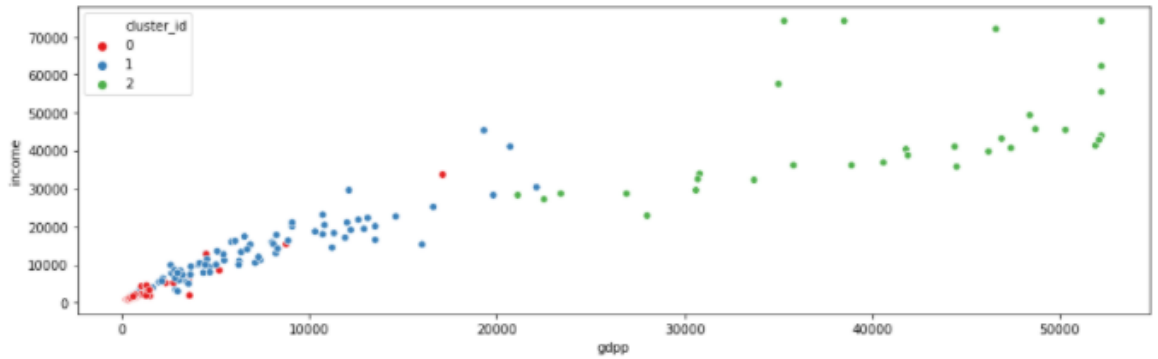


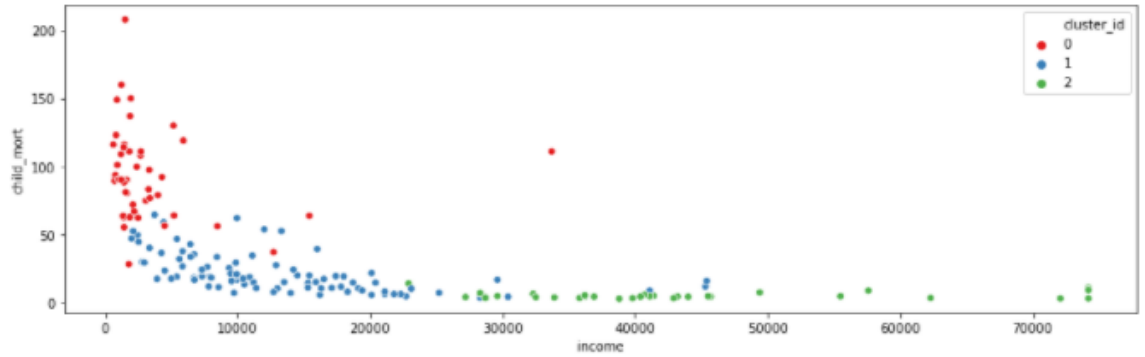


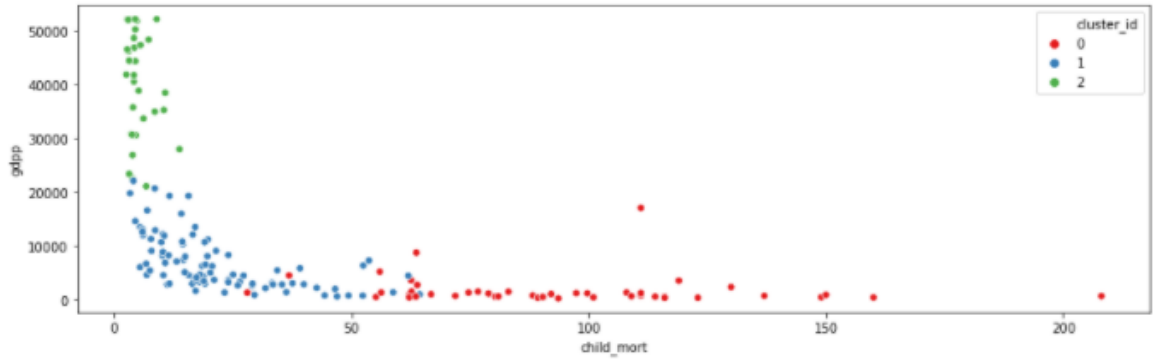
K- means Clustering

- ✓ From the above graphs, We see that 0 shows the signs of lower economic countries and 1 shows the signs of developing countries and 2 shows the signs the developed countries.
- ✓ Distribution between the clusters are uneven

```
1    88
0    46
2    33
Name: cluster_id, dtype: int64
```

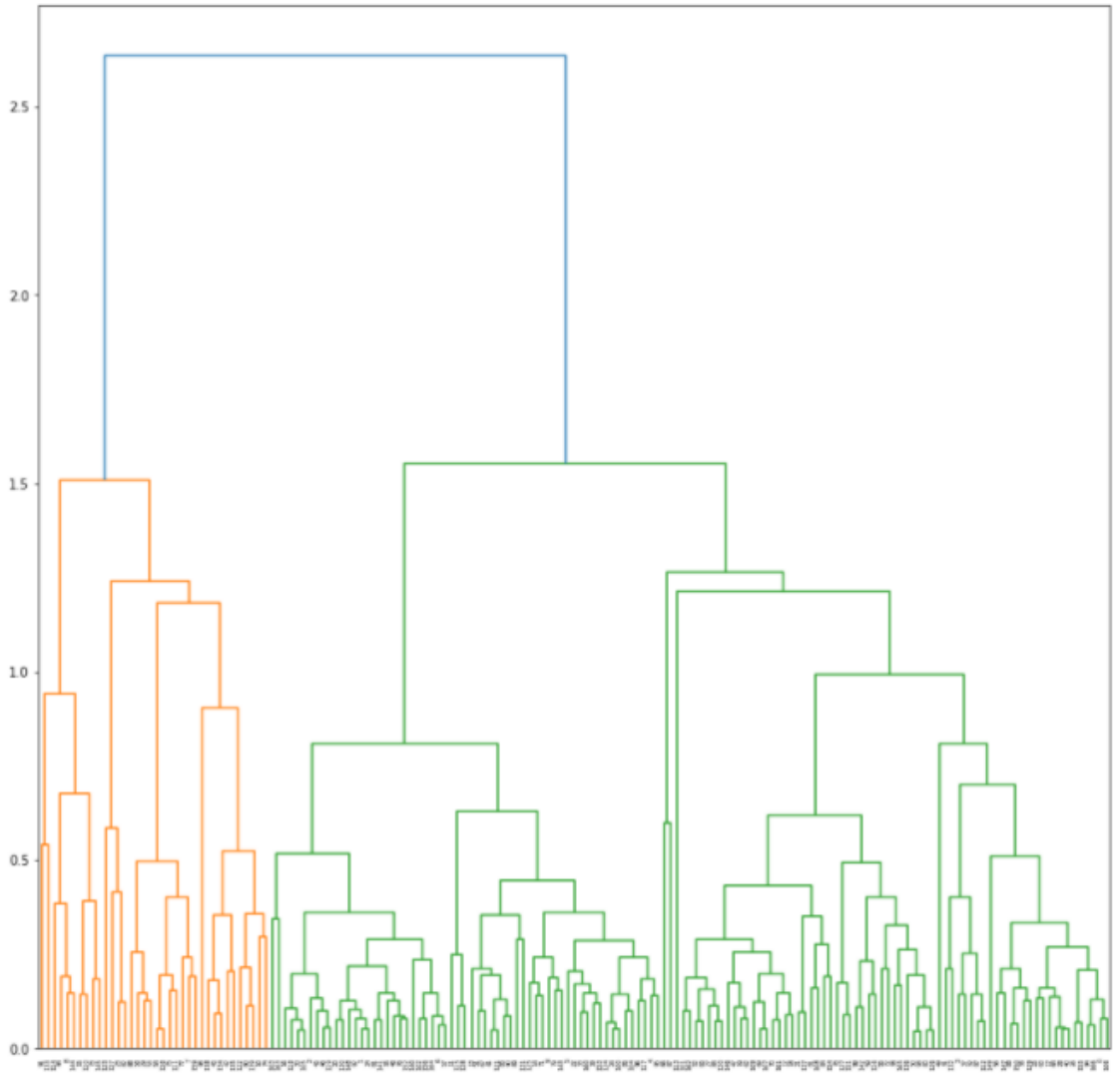






Hierarchical Clustering

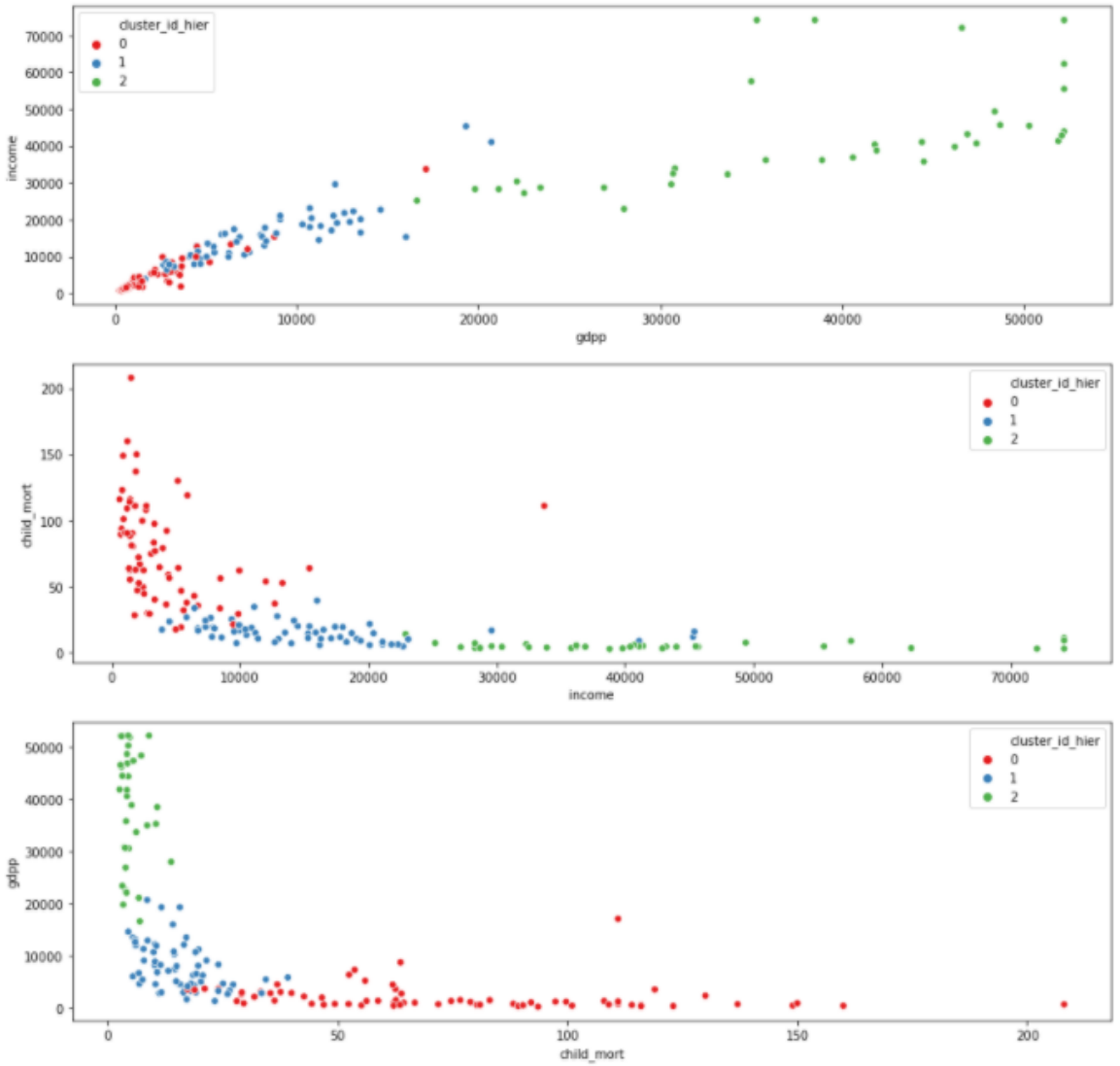
- Complete method has been selected because the cluster is distributed better than the Simple method.
- And cluster has been formed by Cluster K=3



Hierarchical Clustering

- ✓ From the above graphs, We see that 0 shows the signs of lower economic countries and 1 shows the signs of developing countries and 2 shows the signs the developed countries.
- ✓ Distribution between the clusters are uneven

```
0    70
1    61
2    36
Name: cluster_id_hier, dtype: int64
```



Cluster Profiling

From the output, we could label the clusters as

* 0 - Underdeveloped
* 1 - Developing
* 2 - Developed

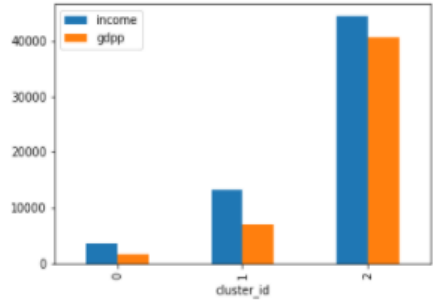
In [150]:
```
# check the characteristics of each cluster
orig_df[['income','gdpp','child_mort','cluster_id']].groupby('cluster_id').mean()
```

Out[150]:

cluster_id	income	gdpp	child_mort
0	3516.804348	1695.913043	93.284783
1	13173.750000	6913.659091	21.923864
2	44437.333333	40726.969697	5.172727

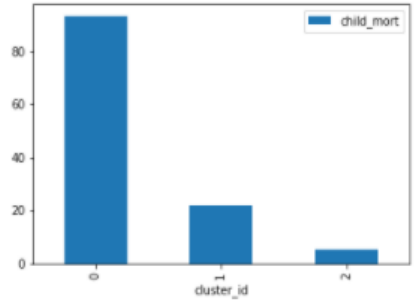
In [151]:
```
orig_df[['income','gdpp','cluster_id']].groupby('cluster_id').mean().plot(kind = 'bar')
```

Out[151]: <matplotlib.axes._subplots.AxesSubplot at 0x25d89057940>



In [152]:
```
orig_df[['child_mort','cluster_id']].groupby('cluster_id').mean().plot(kind = 'bar')
```

Out[152]: <matplotlib.axes._subplots.AxesSubplot at 0x25d88cfb940>



INFERENCES

Countries which require utmost funds based on gdpp, child_mort and income as per hierarchical clustering complete linkage and K means clustering (K=3) are

- * Burundi
- * Liberia
- * Congo, Dem. Rep.
- * Niger
- * Sierra Leone
- * Madagascar
- * Mozambique
- * Central African Republic
- * Malawi
- * Eritrea


THANK YOU