Assignment : Clustering of Countries

-- Seema S B

Problem Statement

- HELP International NGO main purpose is to help the countries which are fighting poverty and economically backward countries with basic amenities and relief.

- They want to use the money raised by funding programmes strategically and effectively to help 5 countries based on socio-economic and health factors.

Approach

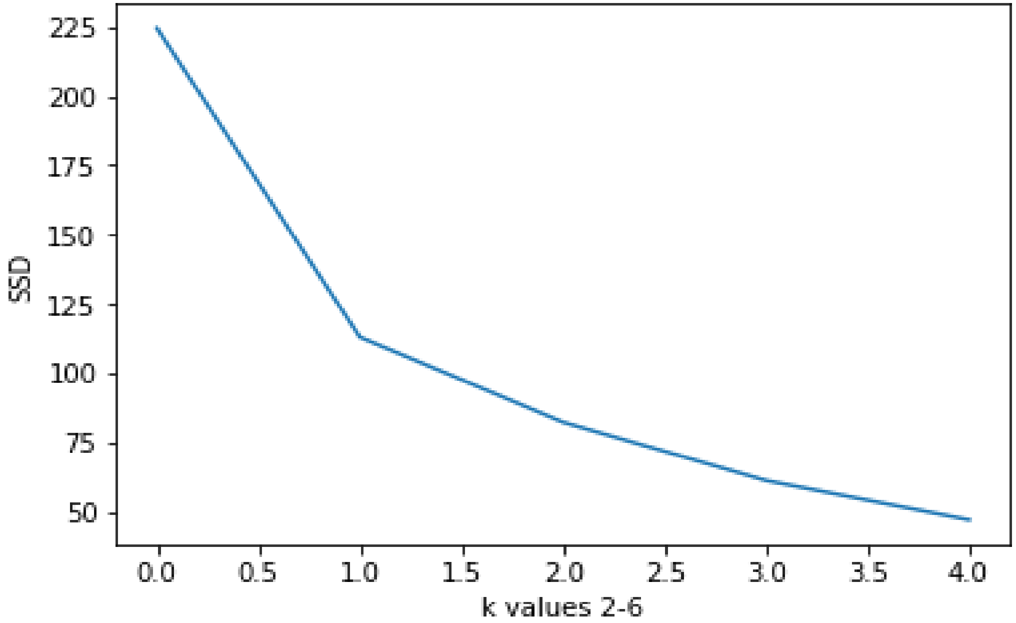
- So, to find the countries based on the socio-economic and health factors, we will see which are the parameters have impact.

- Some of the parameters that tell about a country's status are :
  - The income of each individual in a country
  - The GDP of the country
  - The contribution amount to health sector
  - The child mortality etc

- Here in this problem statement, we are considering the columns GDPP, income and child morality.

Algorithm

- To find the 5 countries based socio-economic and health factors, Clustering is used.
- The clusters obtained by performing K-Means and Hierarchical Clustering are similar.
- The outliers were treated by capping them as the concentration is on the countries with low income and gdpp and high child mortality.
- Based on Elbow curve and Silhouette score, K value is chosen to be 4. Please find the Elbow curve below,



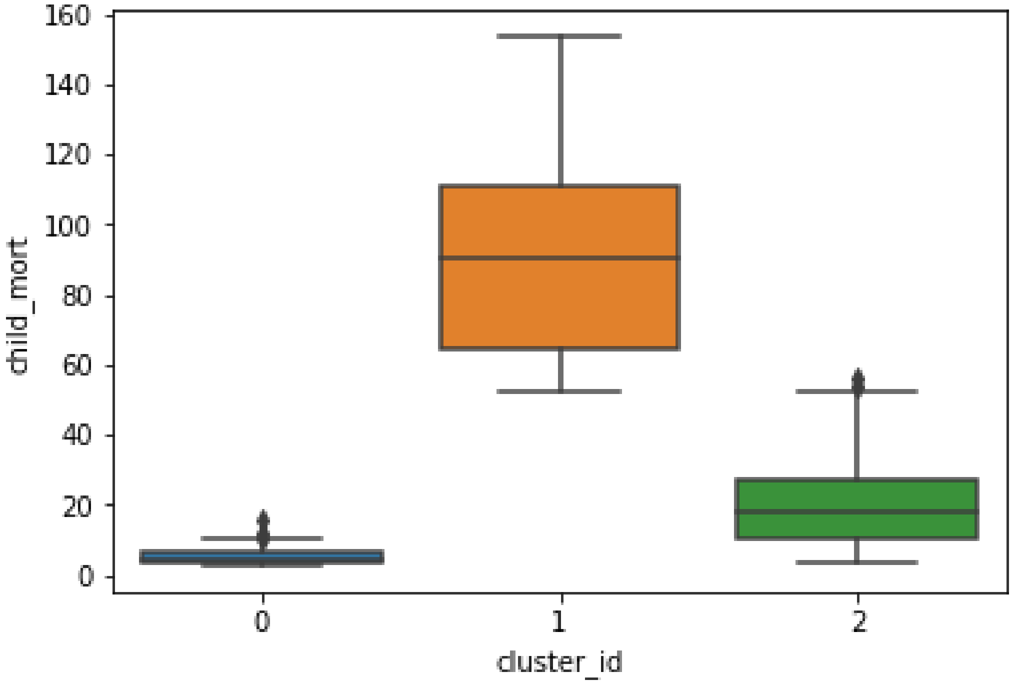
- The Silhouette score are for different K values,

  For n clusters = 2, the silhouette score : 0.566490989486534
  For n clusters = 3, the silhouette score : 0.5376288785294654
  For n clusters = 4, the silhouette score : 0.5366883931729207
  For n clusters = 5, the silhouette score : 0.47921652959750805
  For n clusters = 6, the silhouette score : 0.4641982564829662

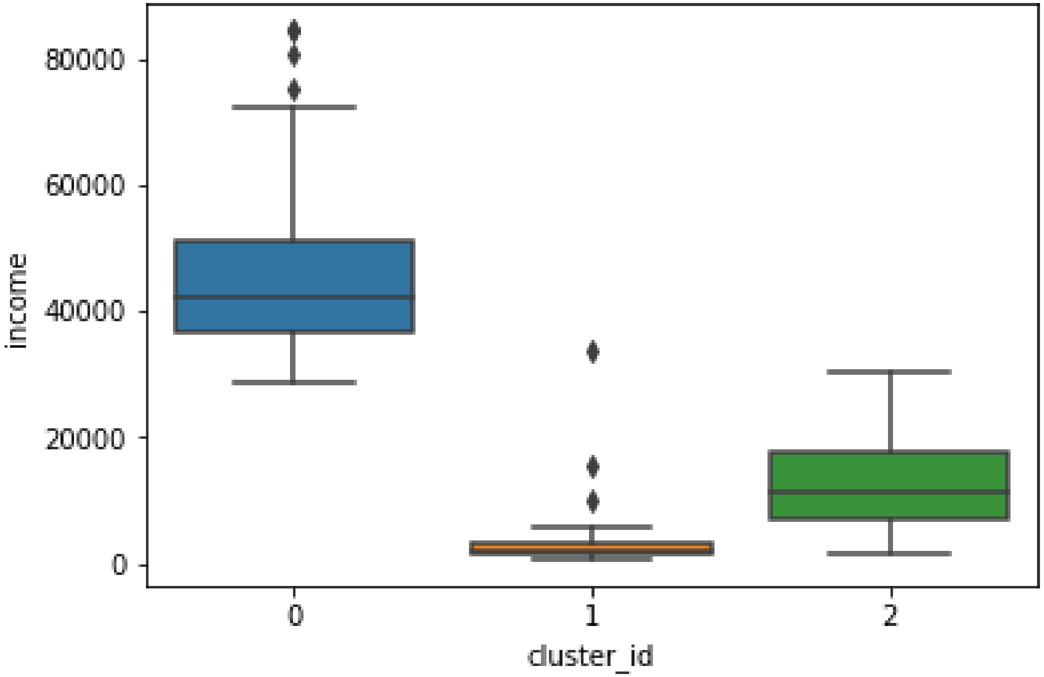
- As the score for K = 3 and K = 4 are almost same 0.54 and there is a significant bent at K=3 in Elbow curve, K is chosen to be 3

- Three distinct clusters are formed. Using which we can clearly say that one of the represents under developed, other developing and another developed countries.

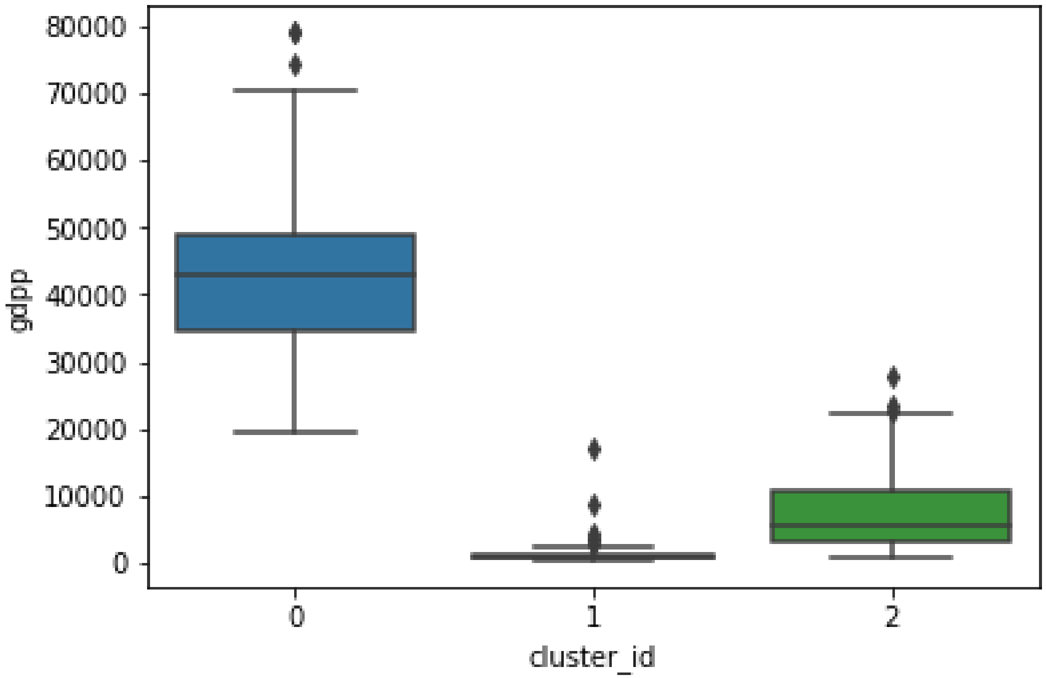
- Below is the graph shows that cluster 1 has highest child mortality rate.



- Below is the graph shows that cluster 1 has low income.

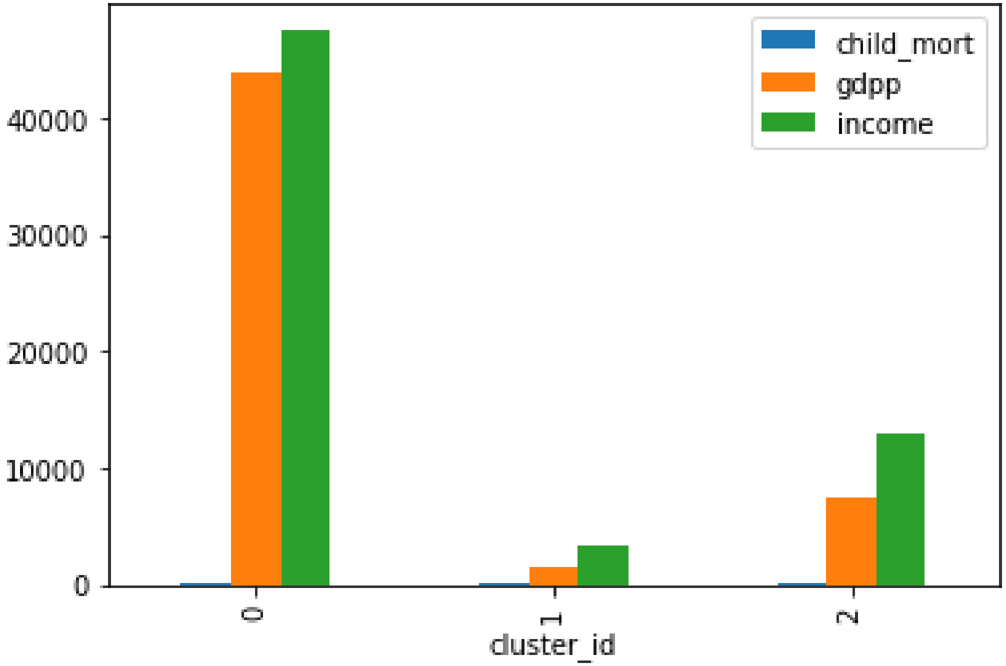


- Below is the graph shows that cluster 1 has low GDPP.



- Below graph and table shows the income, gdpp and child mortality of all clusters, which gives a clear view on which cluster to concentrate.

- Based on which we can say that cluster 1 represents under developed countries as they have low income and gdpp and high child mortality



```
            child_mort          gdpp        income
cluster_id
         0    5.600000  44008.625000  47464.000000
         1   92.478723   1588.366809   3386.988936
         2   20.504545   7358.704545  12924.431818
```

- The final list of 5 countries which are in direst need of aid are
  - Liberia
  - Burundi
  - Congo, Dem. Rep.
  - Niger
  - Sierra Leone

- These countries are selected from the cluster of under developed countries having lowest GDPP.

- GDPP is considered to decide final list of 5 countries as GDPP gives an overall view of the country's socio-economic status and also its contribution for health sector.

- From the given data, countries having low GDPP have low income per person. This also implies countries have less funds to allocate towards health sector. So it can be said that child mortality would be high.