CLUSTERING COUNTRIES ASSIGNMENT

Understanding the Business Problem

Data collected here is for 167 countries across the world on following factors which are the primary indices for measuring the growth of the country and on this basis we have clustered the countries. Following were the factors:

1. Imports
2. Exports
3. Income
4. Life expectancy
5. Health
6. Total Fertility
7. Child Mortality
8. Inflation
9. GDPP

ML Technique used/Evaluating the impact of factors

- We have used the Unsupervised Learning method (clustering) to categorize the countries.
- To analyze the data we have taken the help of the data dictionary to understand the meaning of variables and their impact in accessing the growth of countries.
- After understanding the meaning of each variable we can say that if following variables/features are towards an increasing trend then the country is developed:

1. Imports
2. Exports
3. Income
4. Health
5. Life expectancy
6. GDPP

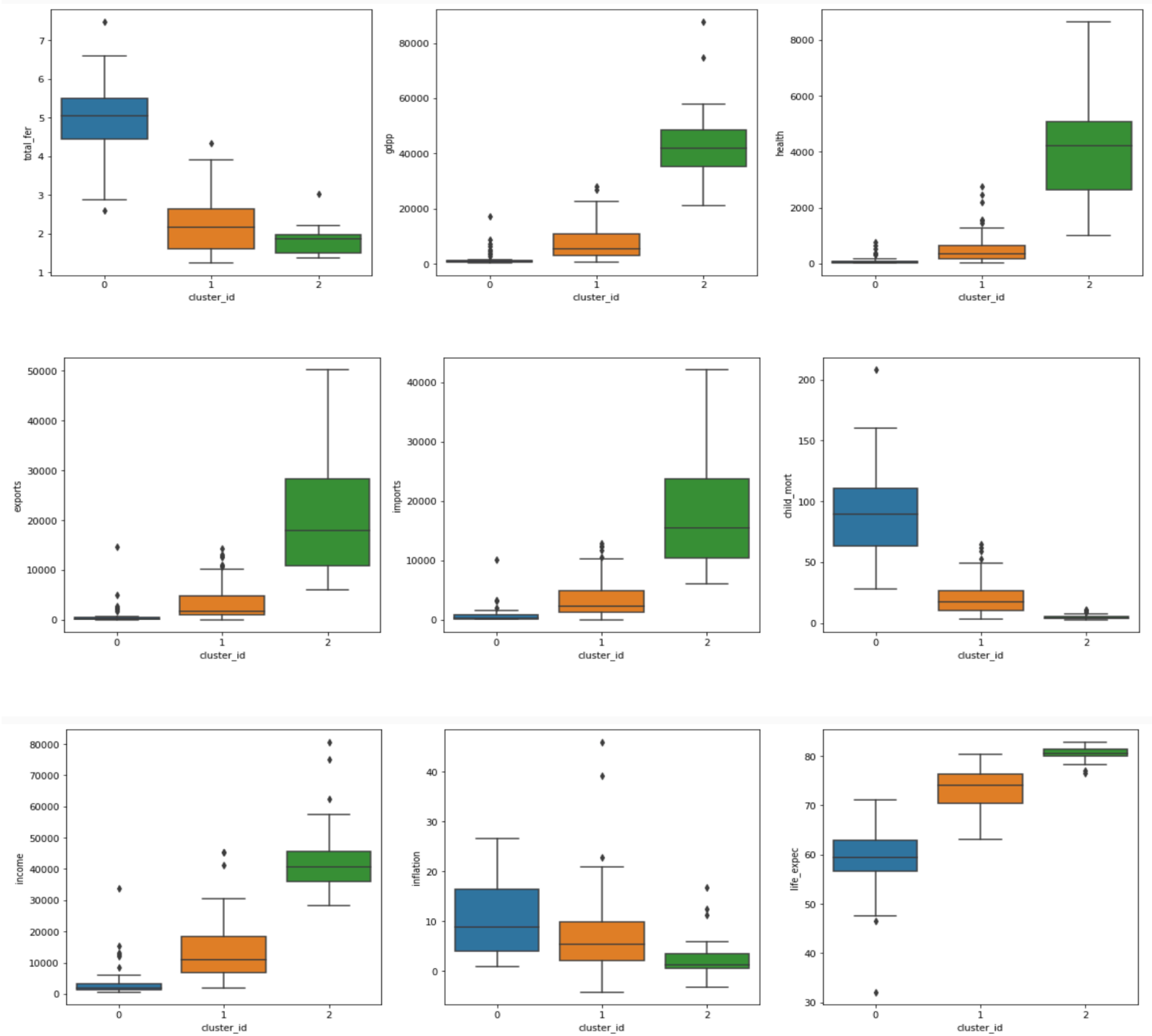
and if the following variables/features are towards an increasing trend then the country is under developed:

1. Total Fertility
2. Inflation
3. Child Mortality

Results of KMeans Clustering:

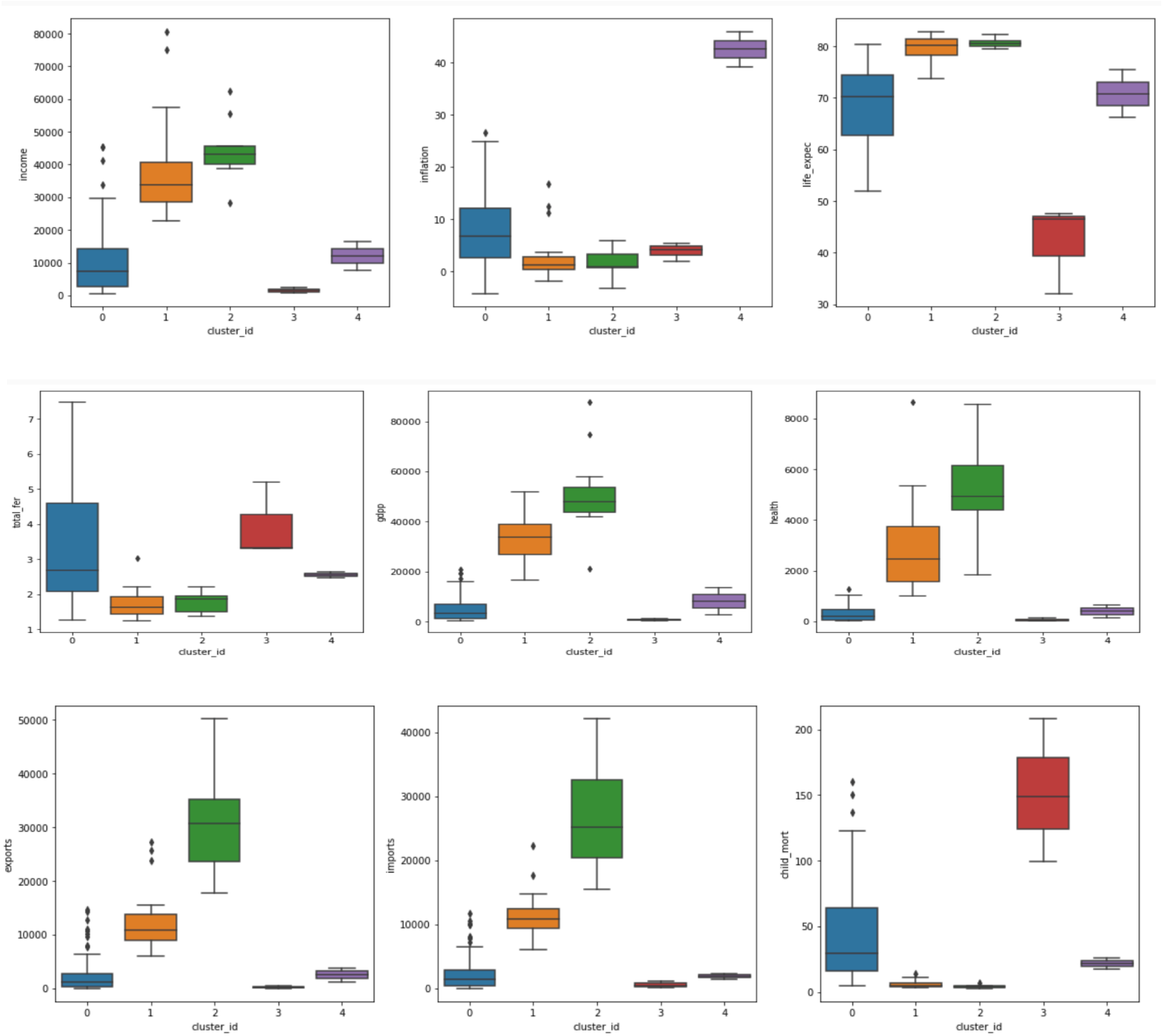
- Results shown here are after performing the KMeans clustering
- The no of optimal clusters obtained are three
- Boxplots of different factors for all the three clusters are shown
- Results can be summarized as:

1. Cluster_0 -- Blue (Under Developed)
2. Cluster_1 -- Orange (Developing)
3. Cluster_2 -- Green(Developed)



Results of Hierarchical Clustering:

- Results shown here are after performing the hierarchical clustering
- The no of optimal clusters obtained are five
- Boxplots of different factors for all the three clusters are shown
- Results can be summarized as:
  1. Cluster_0 -- Blue (Under Developed)
  2. Cluster_1 -- Orange (Developing)
  3. Cluster_2 -- Green(Developed)
  4. Cluster_3-- Red (Under Developed)
  5. Cluster_4 -- Purple (Under Developed)



Categorizing countries on the basis of clusters formed

- We have scaled the data, then used following two clustering methods to group countries
- After the clusters labels have been given to each country we have made a box plot for all the clusters to understand which type of country fall into which category

K-Means:

Here the no of the clusters is three and respective count of the countries is as follow:

1. Cluster_0 47 Under Developed
2. Cluster_1 89 Developing
3. Cluster_2 27 Developed

Hierarchical:

Here the no of the clusters is five and respective count of the countries is as follow:

1. Cluster_0 125 Under Developed
2. Cluster_1 21 Developing
3. Cluster_2 12 Developed
4. Cluster_3 3 Under Developed
5. Cluster_4 2 Under Developed

Shortlisting the countries

- Results obtained in terms of the no of clusters is different, however the countries which need help from the organization have been shortlisted on the following factors:
  - Child mortality
  - Income
  - GDPP
- As per the K-Means and Hierarchical clustering results five countries which are in dire need to help are :
    - Haiti
    - Sierra Leone
    - Chad
    - Central African Republic
    - Mali