**Clustering the Countries by using Unsupervised Learning for HELP International**

**Objective:**

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

**About organization:**

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.

**Problem Statement:**

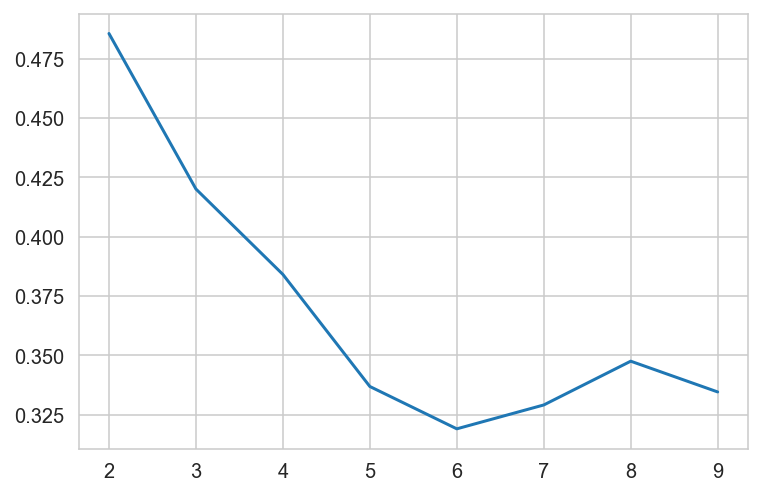
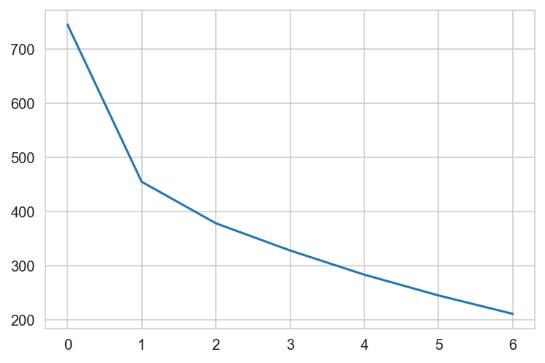
HELP International have been able to raise around $ 10 million. Now the CEO of the NGO needs to decide how to use this money strategically and effectively. So, CEO has to make decision to choose the countries that are in the direst need of aid.

Analysis Approach

**K-Means Clustering**

* **Hopkins Test:**
  + Hopkins score - 0.8421891913312956
  + Since Hopkins score is more than 0.7, we can say that our data have clustering tendency.
* **Optimal value of clusters:**

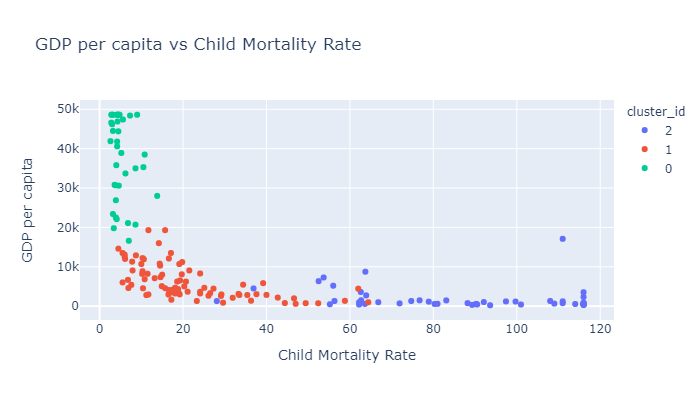
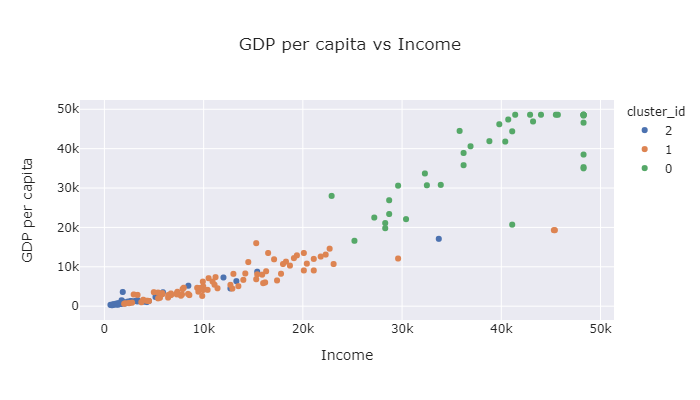
**Elbow curve** **silhouette analysis**

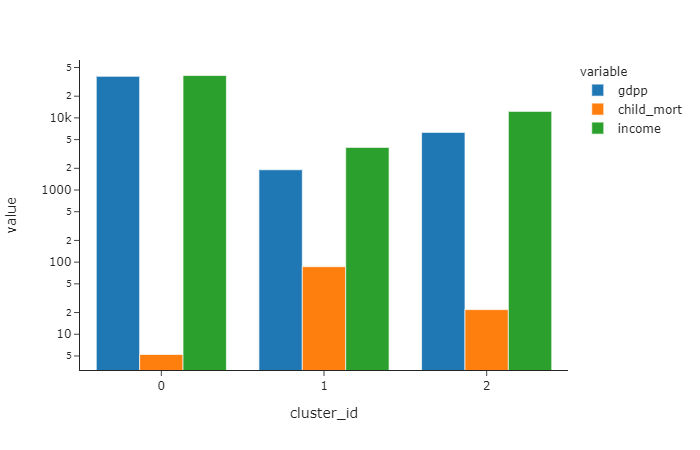
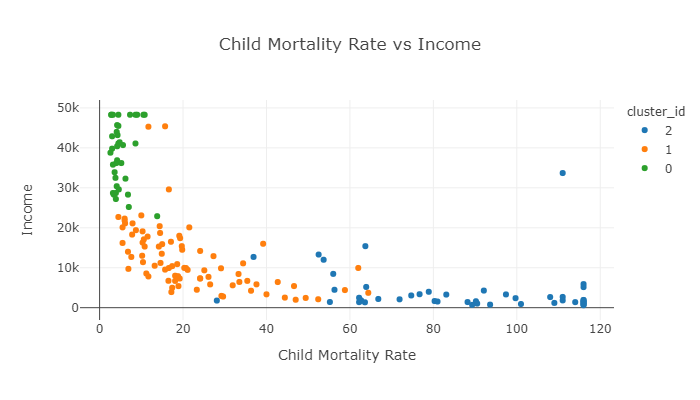


from elbow curve we can see it is suggesting 3 cluster as optimal value since after 3 slope is decreasing at less rate compare to previous 3 cluster. we can see elbow at 3.

average silhouette score for 2 cluster is high, but average silhouette score difference between 2 and 3 is not that big, 3 cluster can also give good quality clusters.

**K-Means Cluster Profiling**

****



**From above graphs we can observe:**

**cluster 2:** countries with moderate level of GDP per capita, moderate level of Child mortality rate and moderate level of Income

**cluster 0:** countries with very high GDP per capita, very low Child Mortality Rate and very high income.

**cluster 1:** countries with very low GDP per capita, very high Child Mortality Rate and very low income.

countries in cluster 0 shows the characteristics Developed Countries & countries in cluster 2 shows the characteristics Developing Countries & countries in cluster 1 shows the characteristics Underdeveloped Countries.

**Cluster 0:** Developed countries

**Cluster 1:** Underdeveloped Countries

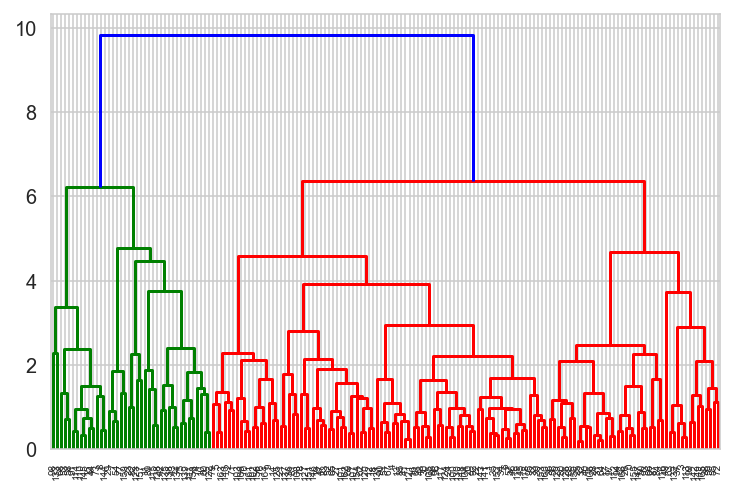
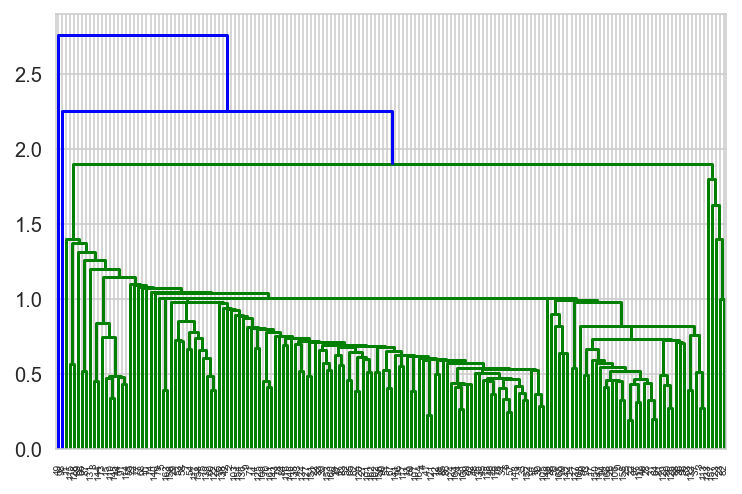
**Cluster 2:** Developing countries

**As per the K-means algorithm, countries which are in direst need are as follow:**

|  |  |
| --- | --- |
| 1. **Burundi** | **6. Madagascar** |
| 1. **Liberia** | **7. Mozambique** |
| 1. **Congo, Dem. Rep.** | **8. Central African Republic** |
| 1. **Niger** | **9. Malawi** |
| 1. **Sierra Leone** | **10. Eritrea** |

**Hierarchical Clustering**

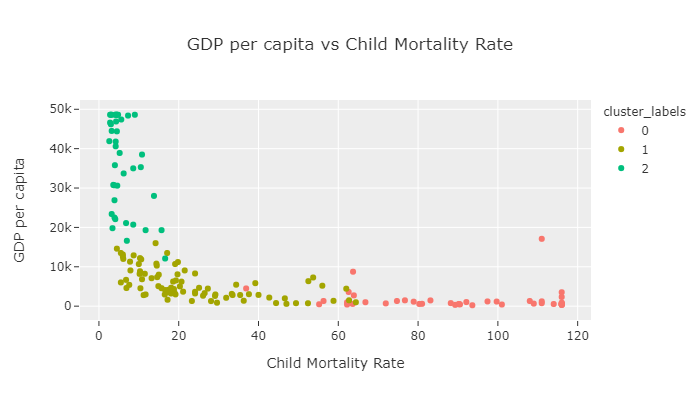
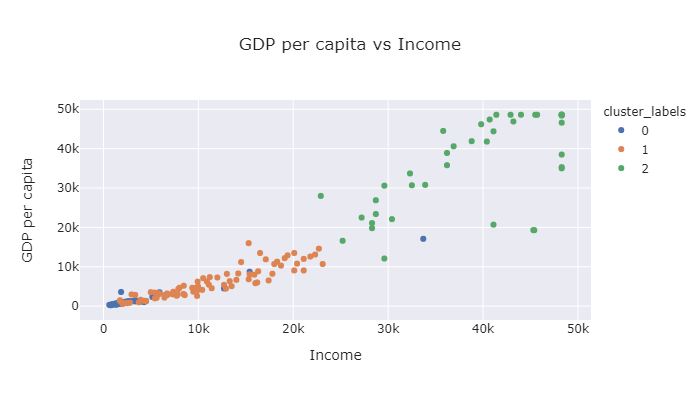
**Single Linkage Dendrogram** **Complete Linkage Dendrogram**

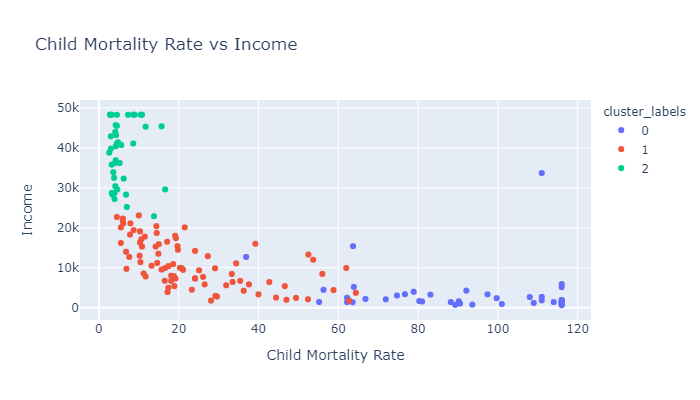
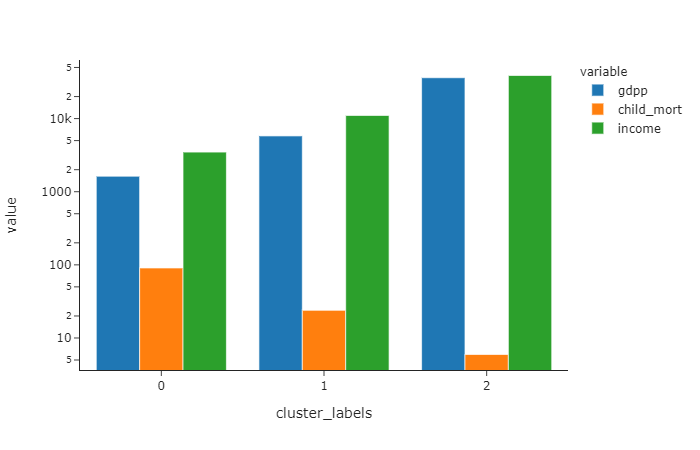


From dendrogram of single linkage and complete linkage, we can see complete linkage clearly have done a great job.

dendrogram with complete linkage suggesting for 2 cluster, but 3 cluster is making more sense to divide countries in 3 groups, to find most needy countries.

**Hierarchical Cluster profiling**

****

**From above graphs we can observe that:**

**cluster 1:** countries with moderate level of GDP per capita, moderate level of Child mortality rate and moderate level of Income

**cluster 2:** countries with very high GDP per capita, very low Child Mortality Rate and very high income.

**cluster 0:** countries with very low GDP per capita, very high Child Mortality Rate and very low income.

countries in cluster 0 shows the characteristics Developed Countries & countries in cluster 2 shows the characteristics Developing Countries & countries in cluster 1 shows the characteristics Underdeveloped Countries.

**Cluster 2:** Developed countries

**Cluster 0:** Underdeveloped Countries

**Cluster 1:** Developing countries

**As per the Hierarchical Clustering algorithm, countries which are in direst need are as follow:**

|  |  |
| --- | --- |
| 1. **Burundi** | **6. Madagascar** |
| 1. **Liberia** | **7. Mozambique** |
| 1. **Congo, Dem. Rep.** | **8. Central African Republic** |
| 1. **Niger** | **9. Malawi** |
| 1. **Sierra Leone** | **10. Eritrea** |

**Summary**

we choose top 10 countries by using both k-means algorithm and Hierarchical Clustering algorithm and we got exact same result.

here first identify the cluster of underdeveloped countries, then based on GDP per capita, child mortality rate and net income per person we identify 10 countries.

* lowest GDP per capita among underdeveloped countries means, these countries are really poor.
* high child mortality rate means, healthcare structure of these countries is in very bad state, maybe here child under age 5, where its immunity system is not developed yet did not have access to different vaccination.
* low net income per person means, person do not have enough income to spend for its own betterment and since the GDP per capita is also low that’s why government also not able to spend on its people.
* we can also observe that selected countries in the final list, all of them are African countries.

**Final List of Countries**

|  |  |
| --- | --- |
| 1. **Burundi** | **6. Madagascar** |
| 1. **Liberia** | **7. Mozambique** |
| 1. **Congo, Dem. Rep.** | **8. Central African Republic** |
| 1. **Niger** | **9. Malawi** |
| 1. **Sierra Leone** | **10. Eritrea** |