Clustering Analysis

Prepared by - SAURABH SINGH

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 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.

Our job is to categorise the countries using some socio-economic and health factors that determine the overall development of the country and come up with the countries that are in the direst need of aid.

Analysis Approach

The steps to be followed for the analysis are mentioned below::

1. Understanding the data.
2. Visualizing the data.
3. Model Building
4. Analyzing the clusters.
5. Cluster Profiling
6. Reporting the top five countries.

Understanding the data

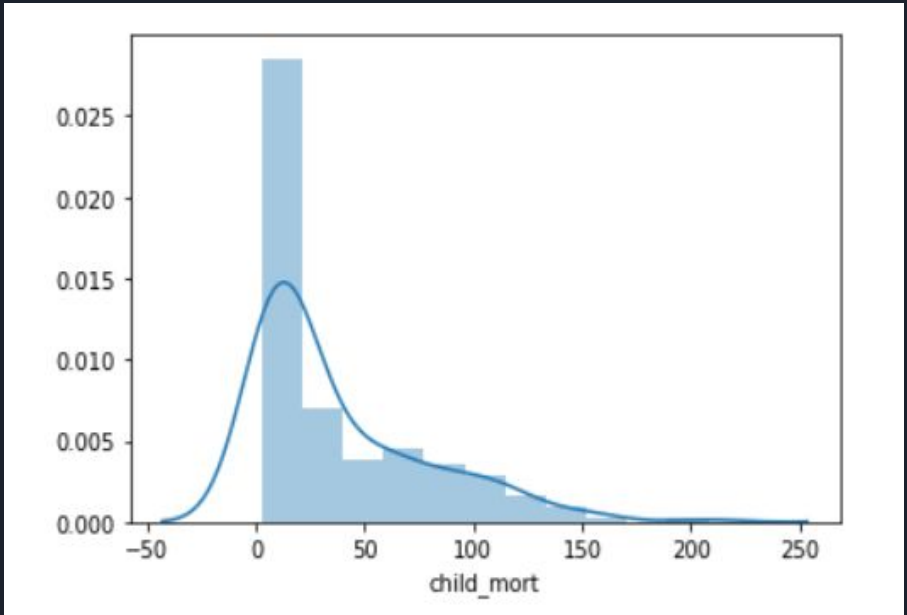
The dataset we are provided with has 167 countries with 10 features.

The features ‘exports’, ‘health’ and ‘imports’ have been provided as the percentages of the GDP per capita. We converted these features back to their original values for better analysis.

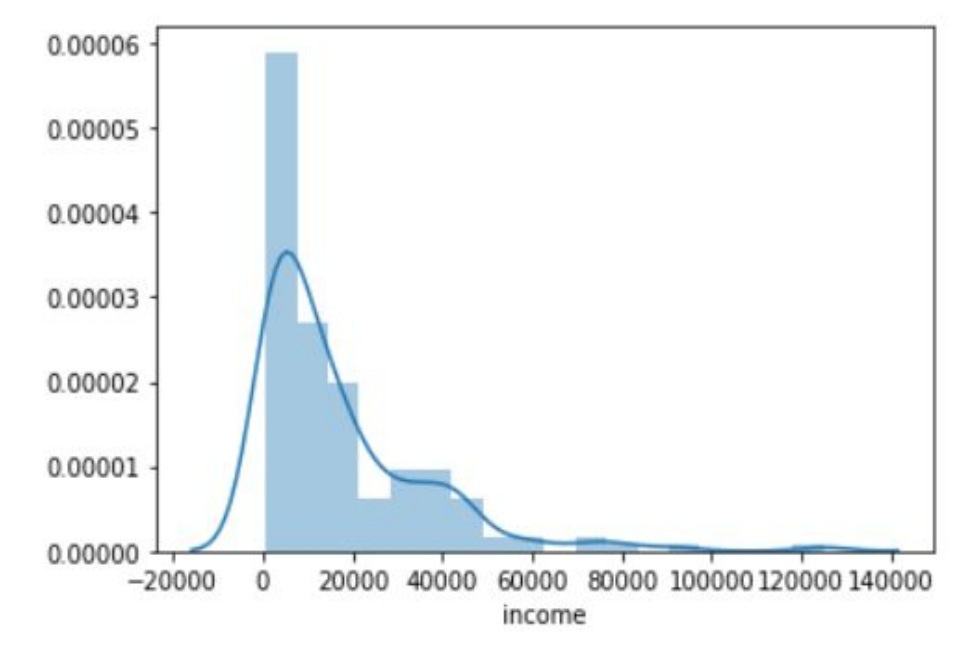
The dataset has no missing values and is in a good and clean format.

Visualizing the data

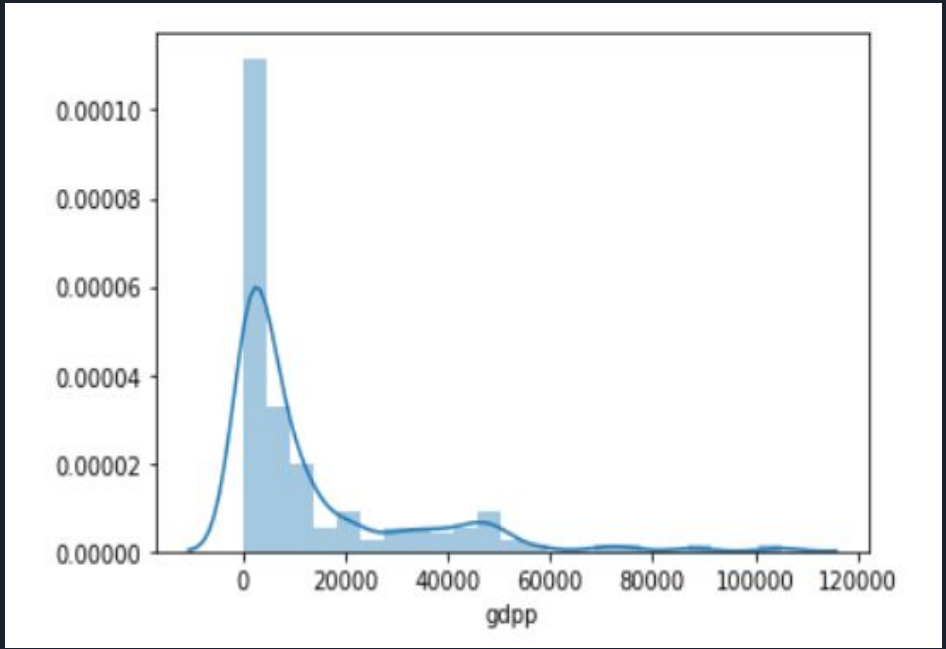
Univariate Analysis



Analyzing the distribution plot of the feature ‘child_mort’, one can infer that though some countries have child mortality above 150, most of the countries have child mortality below 50.

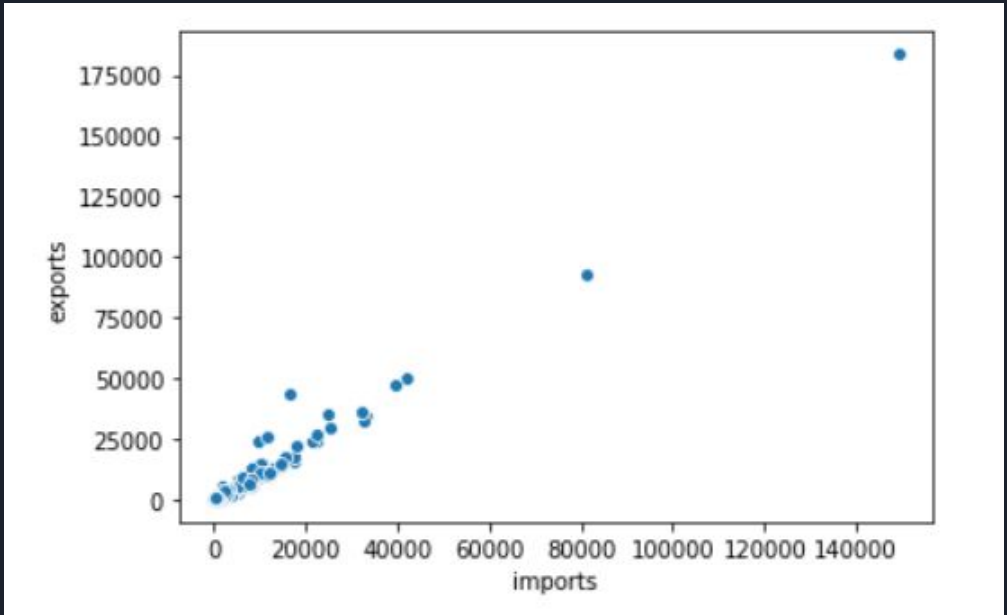


Analyzing the distribution plot of the feature ‘income’, one can infer that the net income per person for most of the countries lies below 20000.

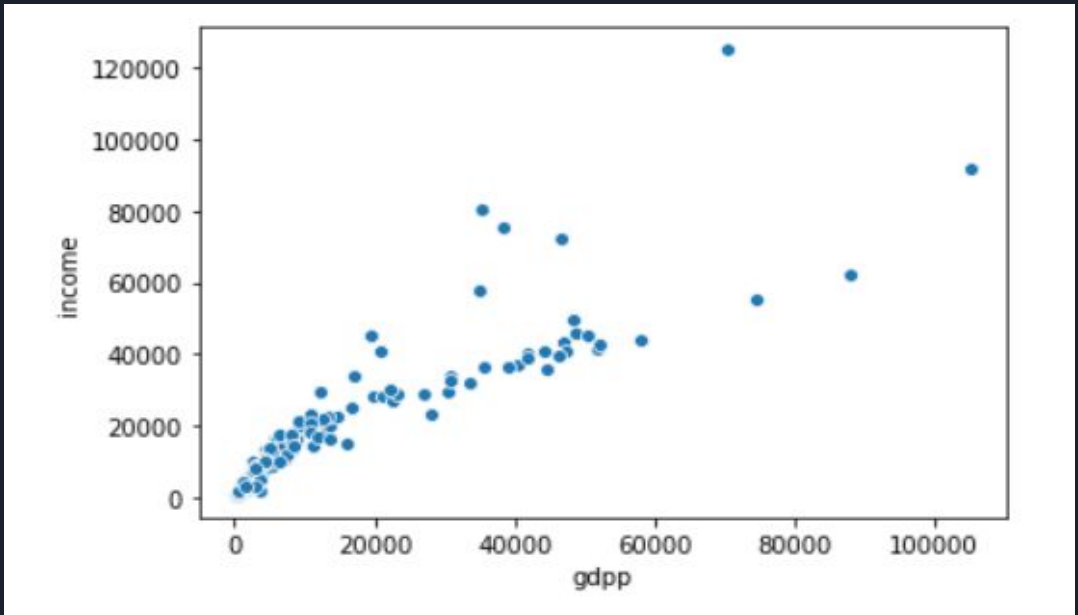


Analyzing the distribution plot of the feature ‘gdpp’, one can infer that the GDP of most of the countries lies below 20000.

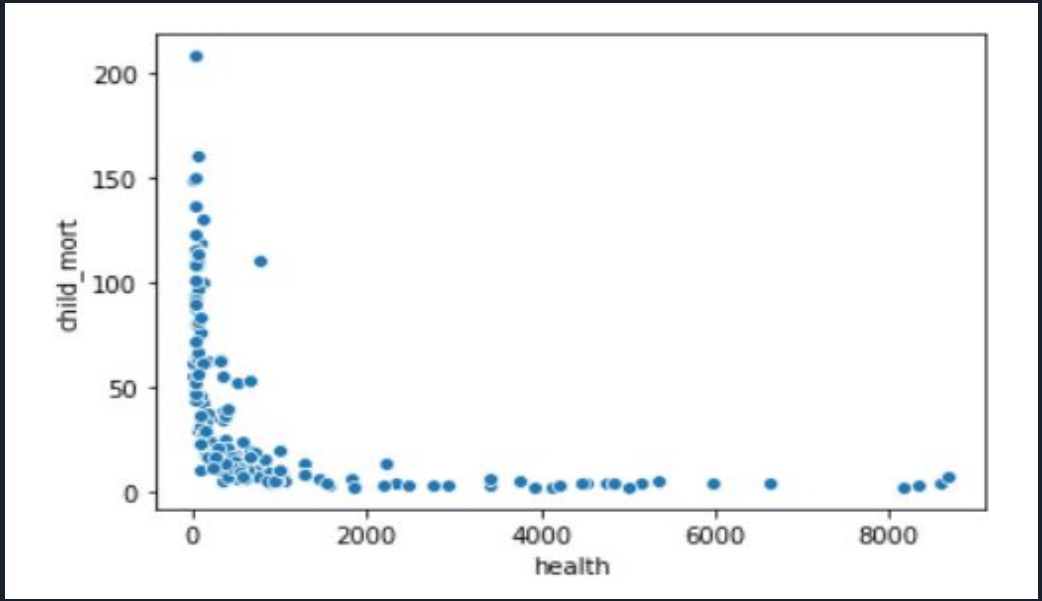
Bivariate Analysis



Couple of countries have exceptional import and export of goods and services per capita, whereas the rest of the countries are concentrated at the bottom left corner of the plot.

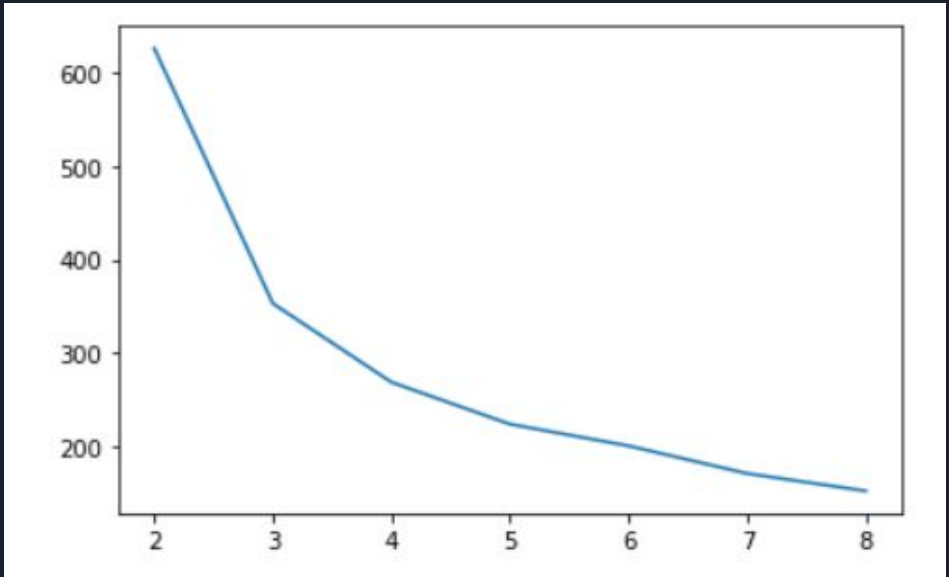


A very few countries have an exceptional income and GDP per capita, whereas the rest of the countries are concentrated at the bottom left corner of the plot.



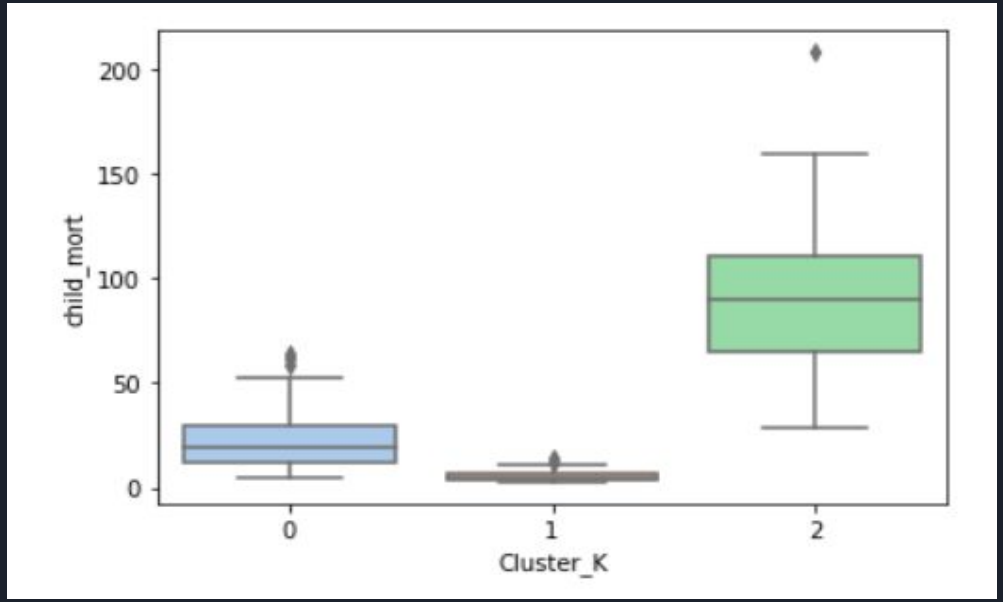
Analyzing the scatter plot, one can infer that the countries with low health spending per capita have a high child mortality rate.

Model Building

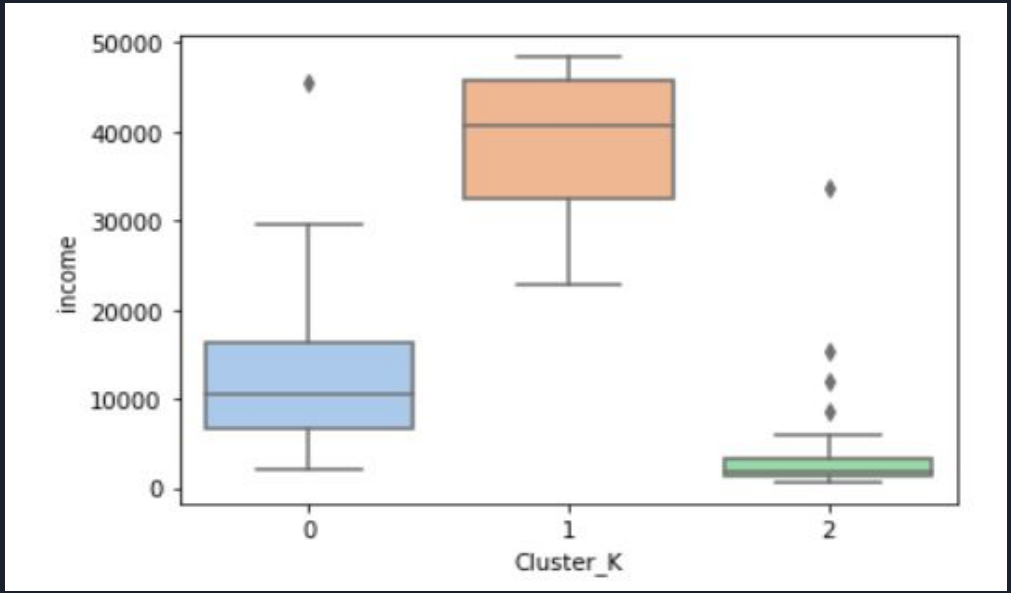


The elbow curve and the silhouette score suggests to cluster the data into 3 clusters.

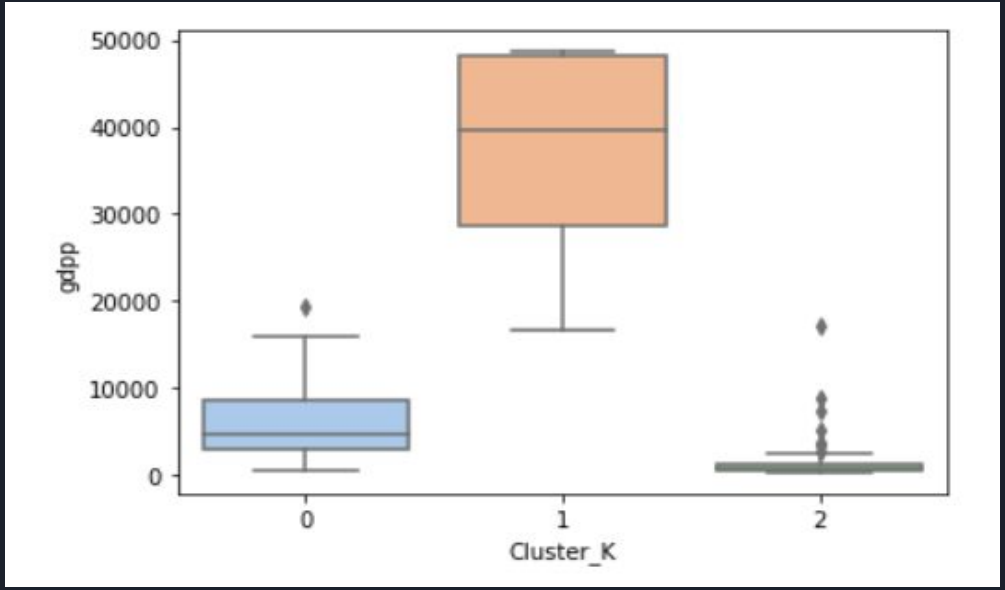
Analyzing the Clusters



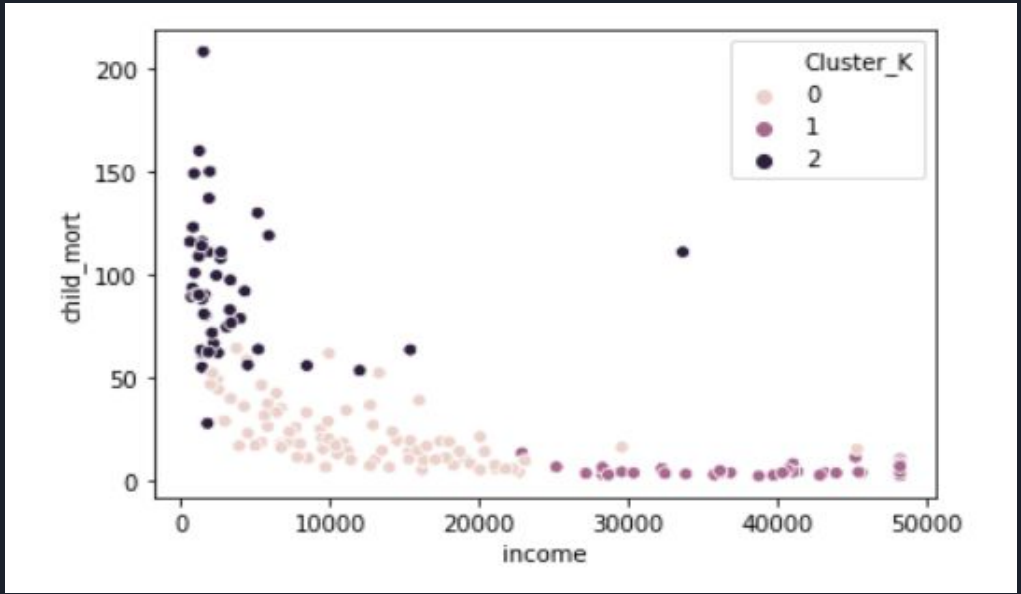
The countries with cluster label 2 have the highest child mortality rate, whereas the countries with cluster label 1 have really low child mortality rate.



The countries with cluster label 2 have net income per person less than 10000, whereas the majority of the countries with cluster label 1 have net income per person greater than 30000.

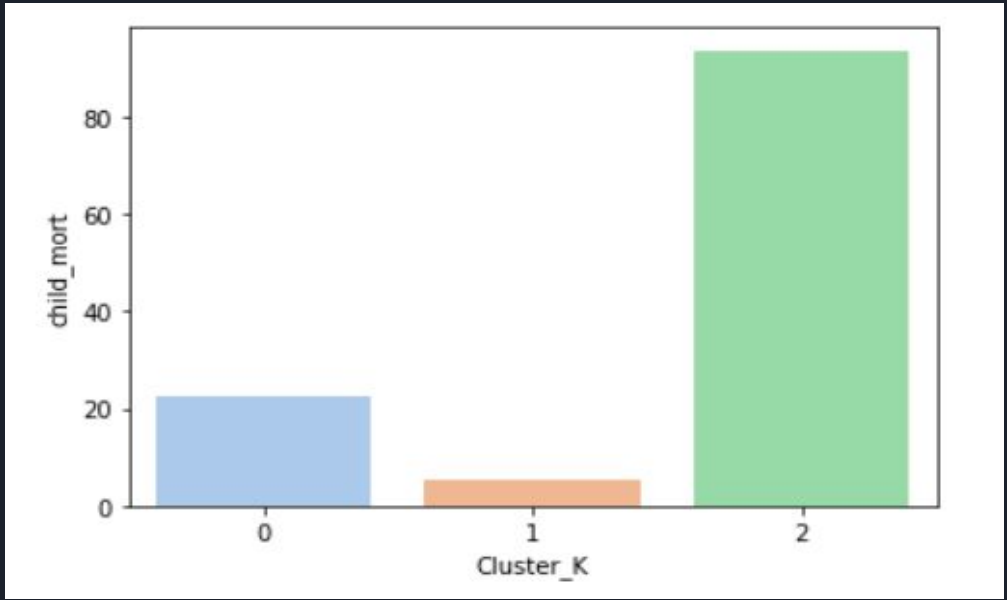


The countries with cluster label 2 have the GDP per capita less than 10000, whereas the majority of the countries with cluster label 1 have the GDP per capita greater than 30000.

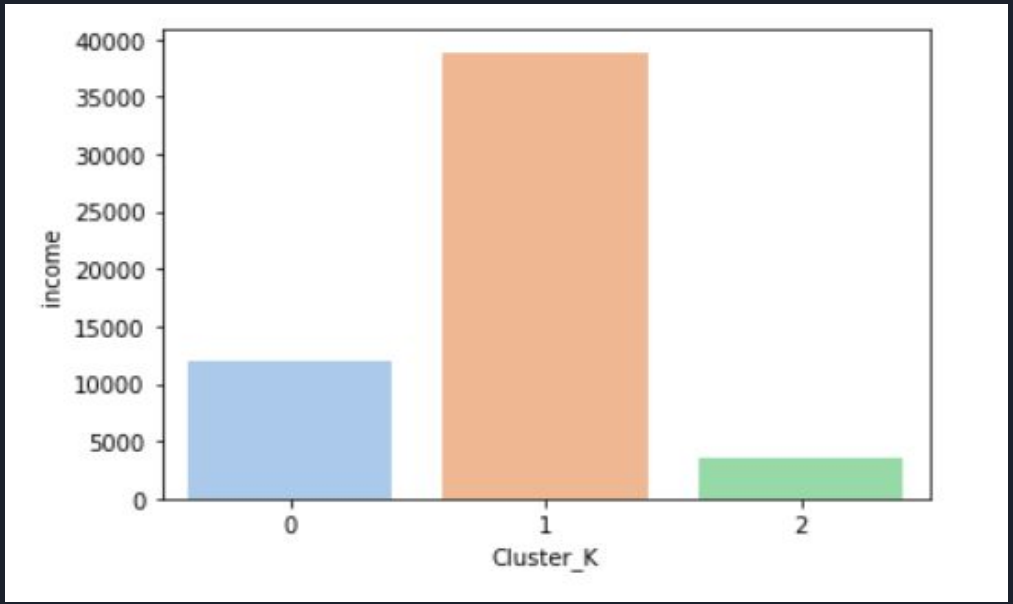


Analyzing the scatter plot between the child mortality and the net income per person, we can clearly see the formation of three clusters.

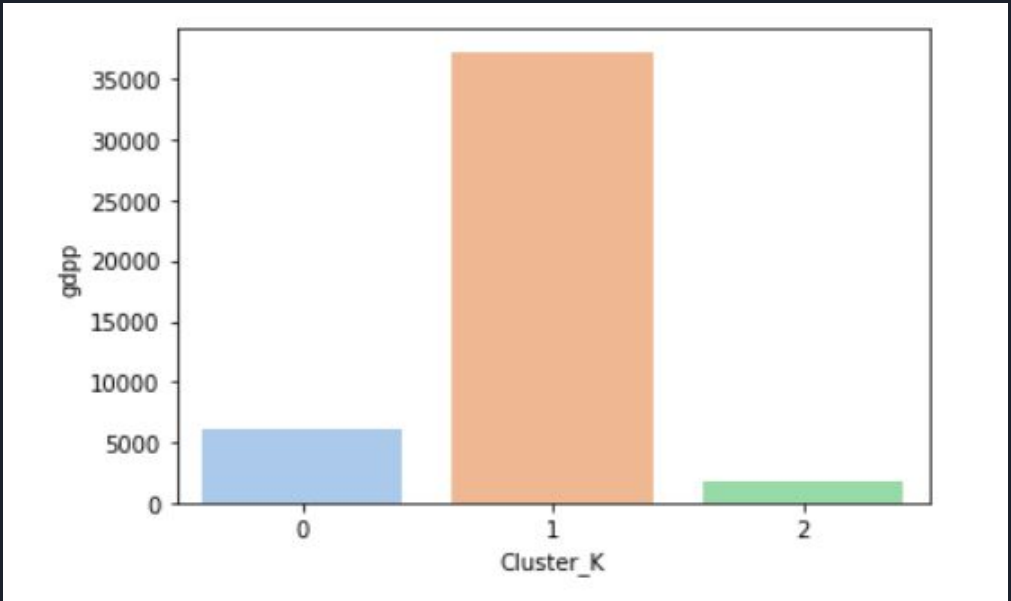
Cluster Profiling



Comparing all the clusters, on an average, we can infer that the countries with cluster label 2 have the highest child mortality rate.



On an average, countries with the cluster label 2 have the lowest income per person.



On an average, countries with the cluster label 2 have the lowest GDP per capita.

The countries that are in the direst need of aid belong to the cluster lanbel 2.

The top five countries that are in the direst need of aid.

1. Haiti
2. Sierra Leone
3. Chad
4. Central African Republic
5. Mali