Minor Project Report

Yuvraj Soni (B21EE089)

- **Abstract :**

HELP International has been able to raise around 10 million Dollars. Now the CEO of the NGO needs to decide how to use this money strategically and effectively. So, the CEO has to make a decision to choose the countries that are in the direst need of aid. In our project We are categorizing the countries using socio-economic and health factors to determine the overall development of the country and help the NGO in helping the needy countries.

- **Dataset** :

The provided dataset contains 167 rows and 10 columns representing 167 countries and 10 socioeconomic and health factors about the countries like child-mortality , exports , health , imports etc.

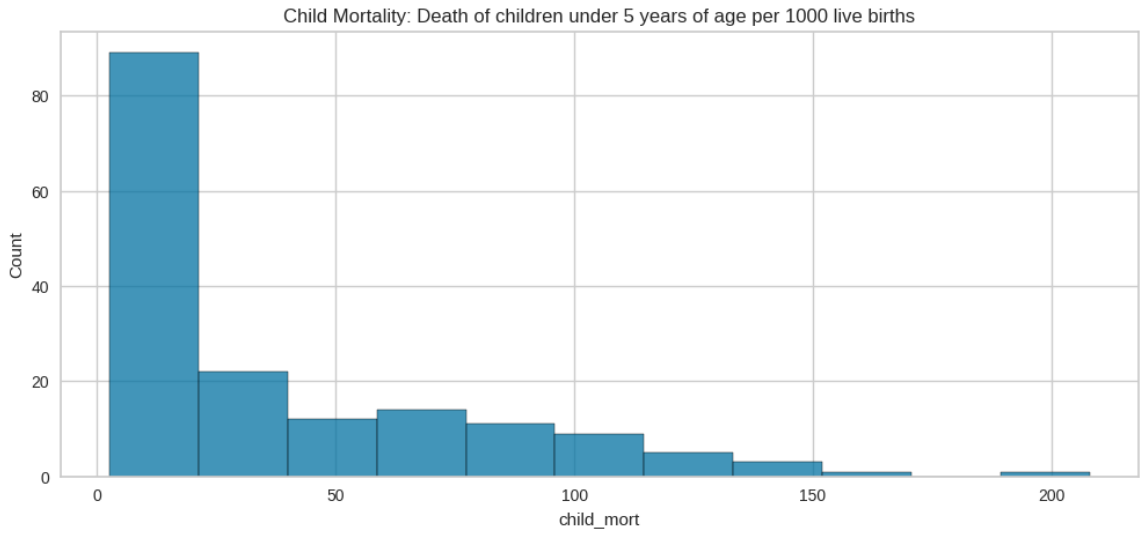
- **Pre-Processing** :

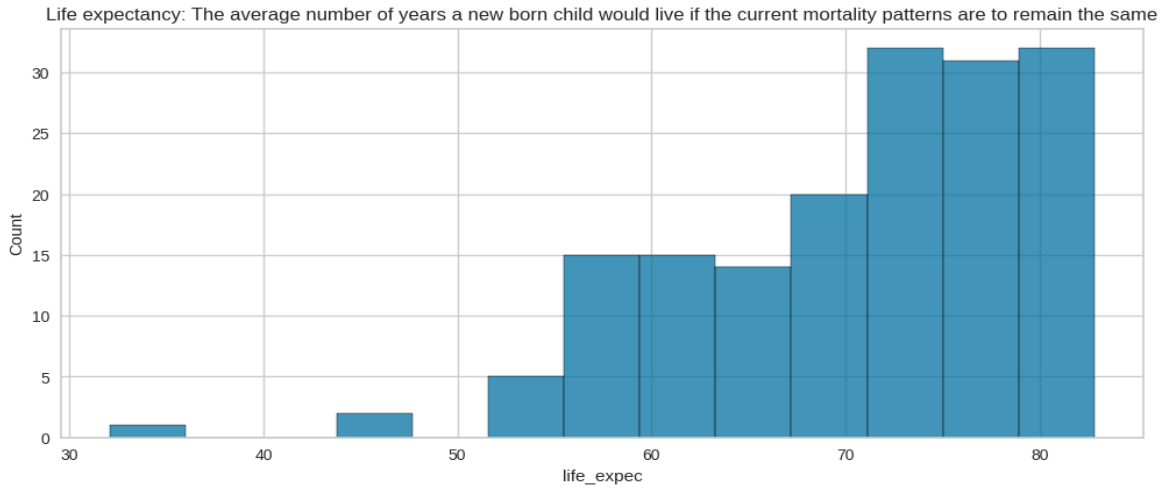
After going through the dataset , we found that the dataset is clean i.e having No missing values and duplicate values and small but having some outliers and

skewed distribution. So it was not required to do so much pre-processing in this dataset.

- **Data-Visualization** :

First of all, we plotted the histplot of all the features of the dataset through which we got a quite good idea about the distribution of the features.Then after we plotted the correlation matrix of the features through heatmap of sns library. In this we found that life expect and total fertility were having very high correlation with child mortality.





We have also plotted the distribution of ‘Exports’,’Imports’,’Health’,’Income’,’Inflation’,’Life Expectancy’ keeping in mind the following points

Features of Economically Backward Countries :

- The country's **per capita income is very low**.
- **High Population** that leads to non - availability of resources.
- **Unemployment** due to less resources.
- **Low country wealth** that leads to **low capital**.
- **Inequitable** distribution of **wealth** and **income**.
- **Lack of proper educational amenities** and thus **illiteracy prevails**.
- **Low level of living**.
- **No technical advancement**.
- **Poor health services** coupled with **high birth & death rates**.

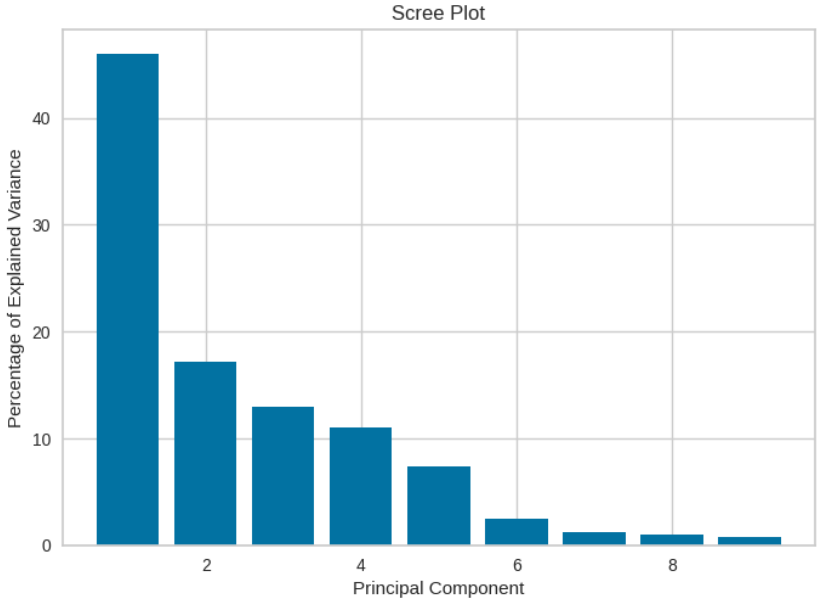
.

- ## Scaling:

We scaled our given dataset using a standard scaler because our dataset was not on the same scale. Also scaling of the dataset to a similar scale can help improve the performance of machine learning algorithms.
Also StandardScaler is less sensitive to outliers than other scaling techniques. Since it is based on the mean and standard deviation, outliers have a smaller effect on the scaling of the data.

- ## PCA:

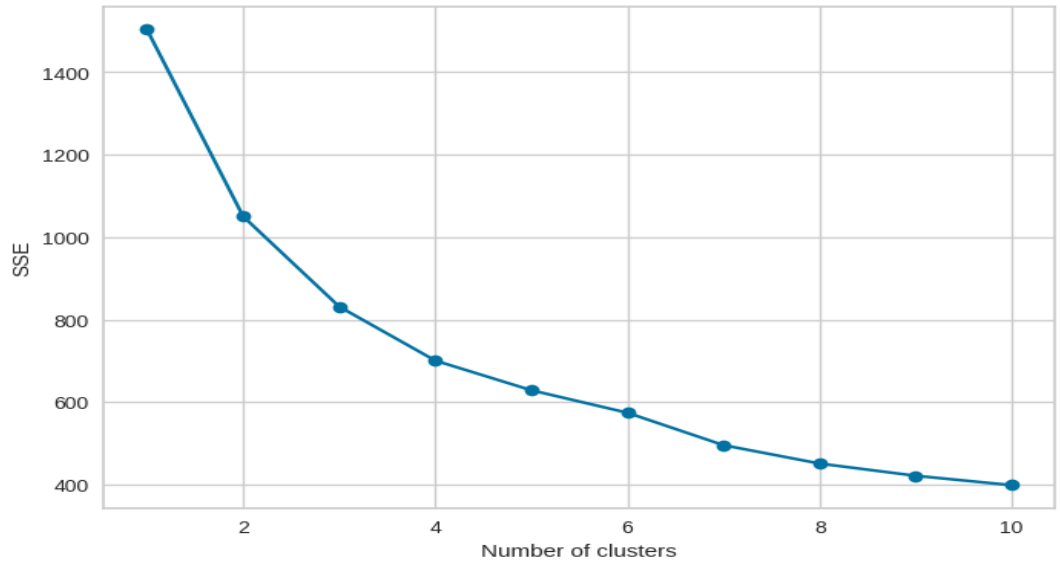
  Applied PCA on the standardised data set and we saw that the the percentage of explained variance crosses 90% with just five components

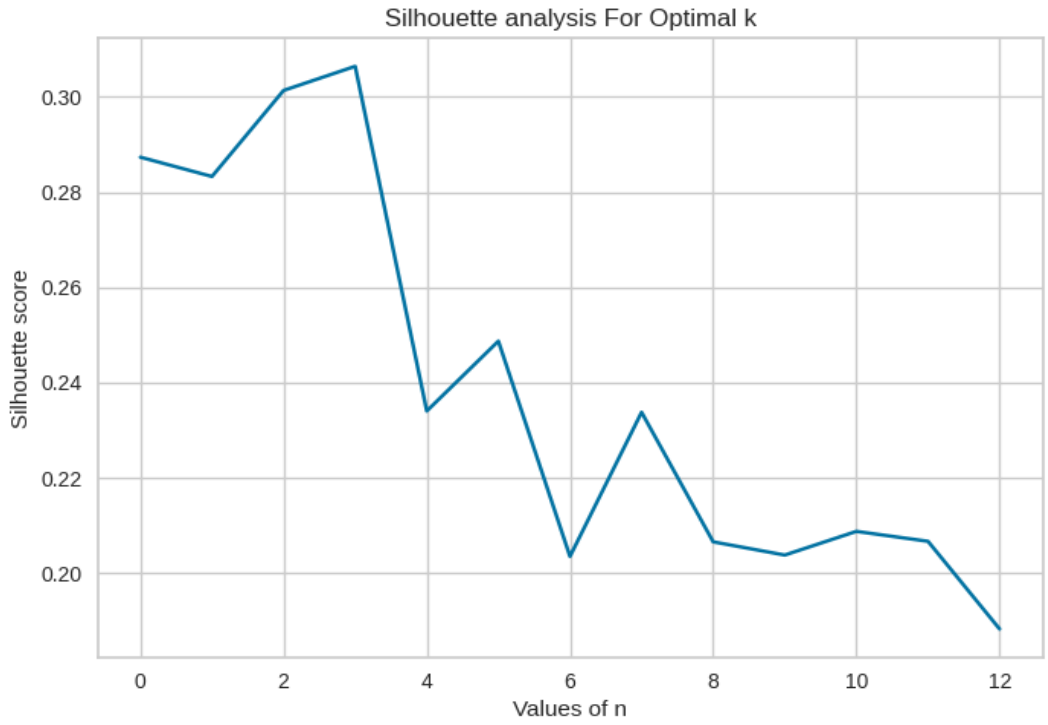


So we reduced the dimensionality of the data set to 5 components.

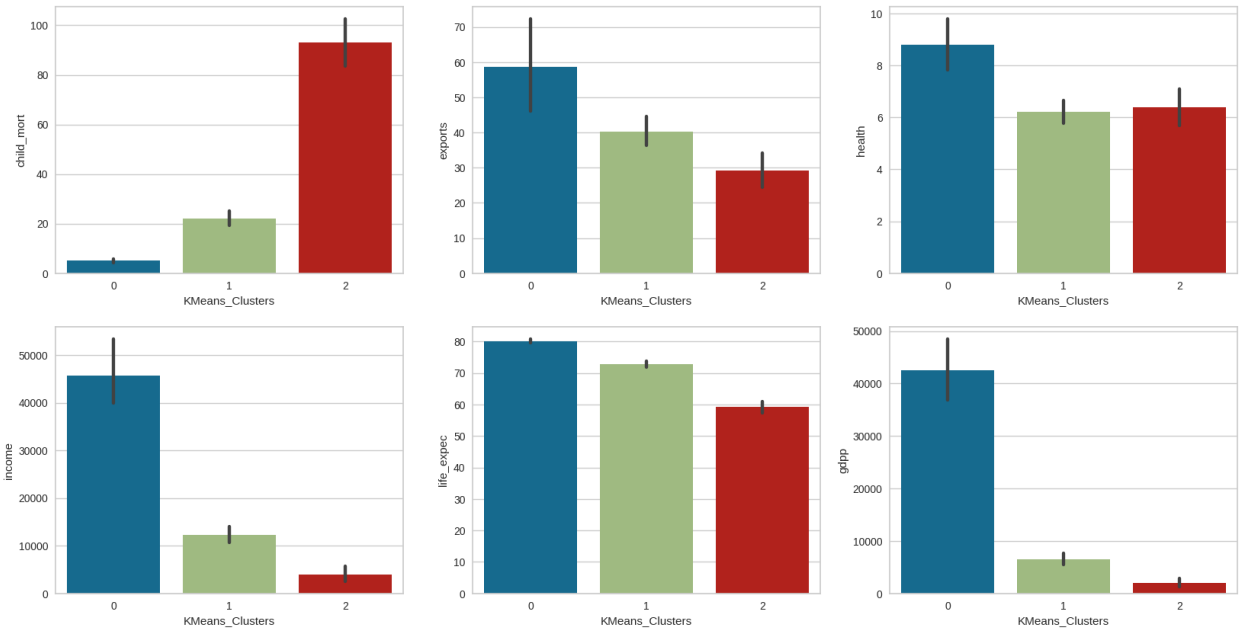
- **K means on normal dataset :**

Firstly we determined the number of clusters to be made on this dataset by Elbow curve and silhouette analysis through which we found that the optimal number of clusters to be formed is 3.

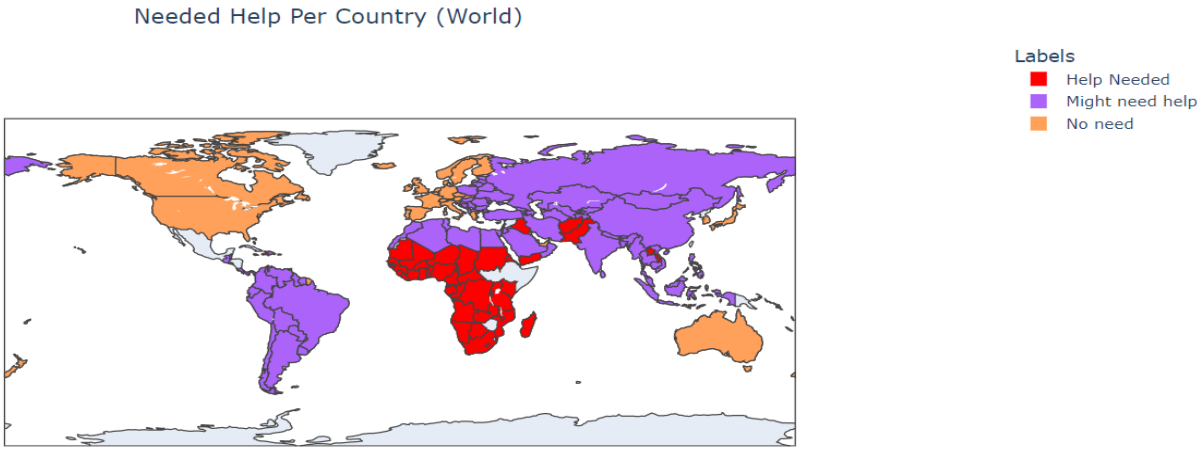




Then I trained the k-means clustering algorithm on standard dataset with the number of clusters 3 .
Then we plotted the scatter plots of the clusters and divided the countries into 3 clusters i.e 0,1,2
From the scatter plots and the bar plots plotted we concluded that the countries that are **clustered 2 are in need of the help** and we named them as **underdeveloped countries**, countries **clustered 1 are named as developing countries** and they might need help, but the **cluster 0** countries have performed poorly according to the criterion set above so they were named as **under-developed countries.**

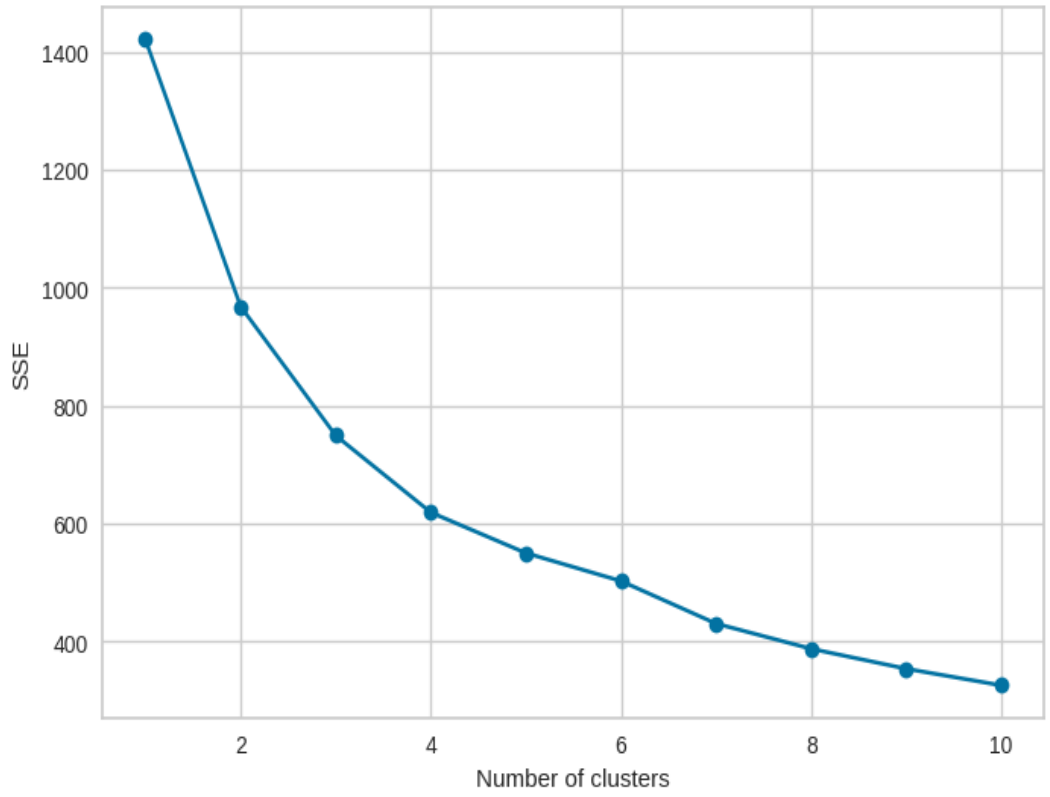


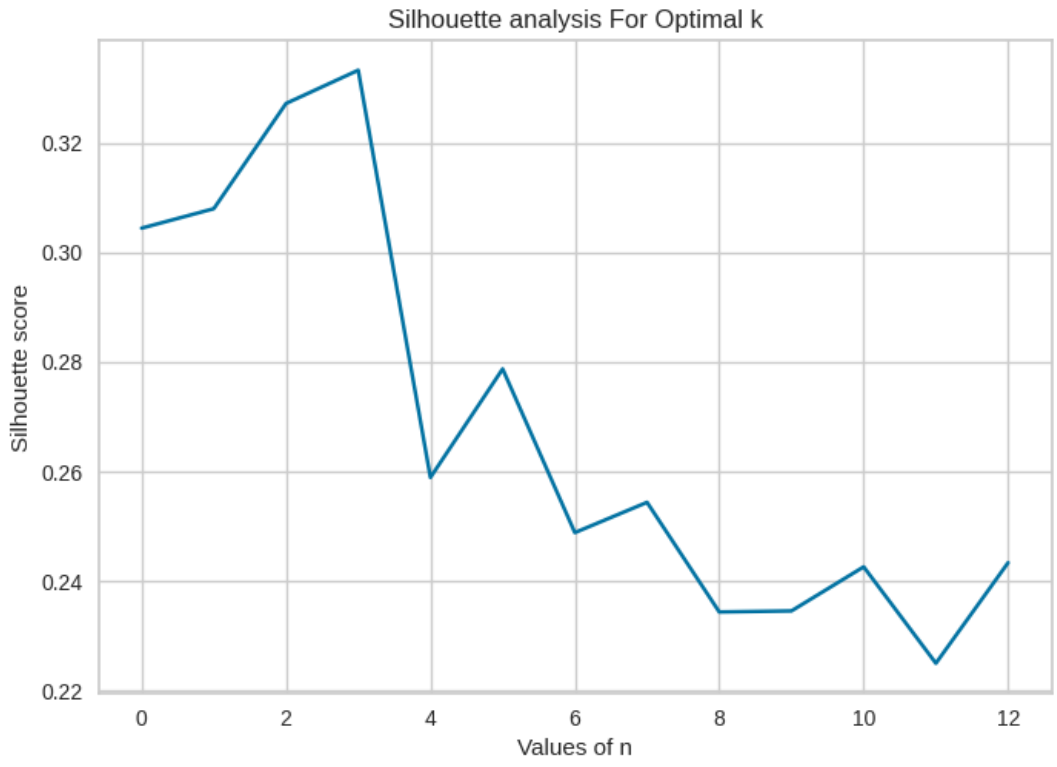
After this we stored the developing , developed and under-developed countries in 3 lists.



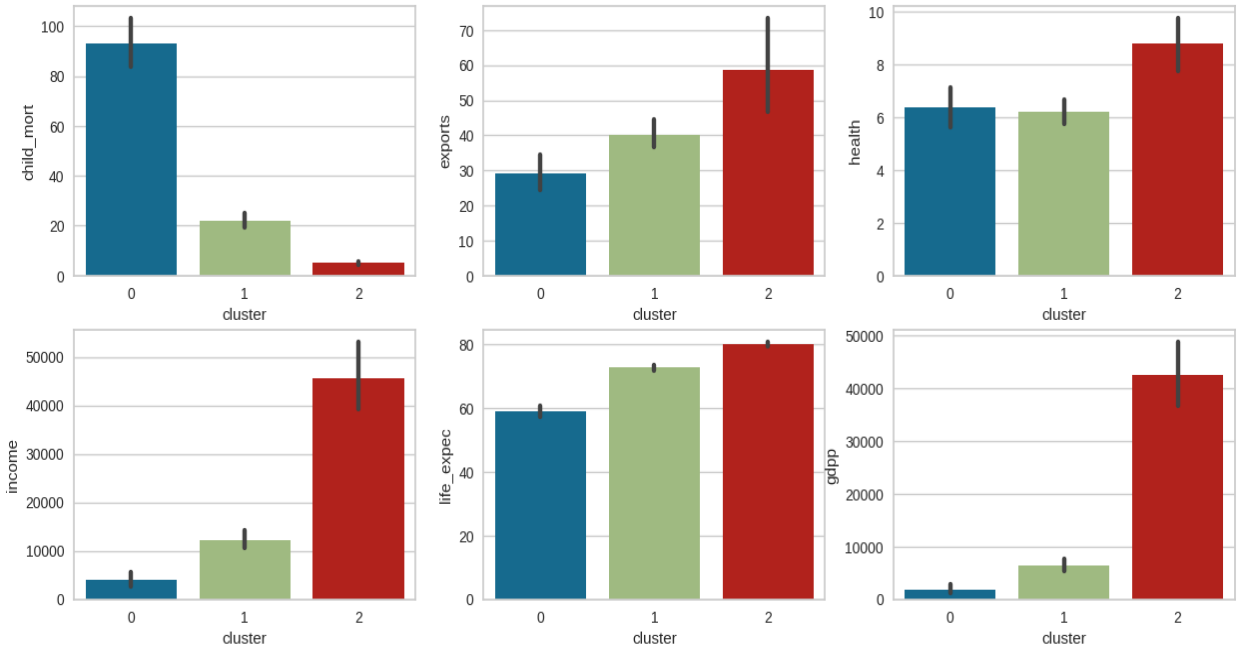
- **K Means on PCA dataset :**

We determined the number of clusters to be made on this dataset by Elbow curve and silhouette analysis through which we found that the optimal number of clusters to be formed is 3.





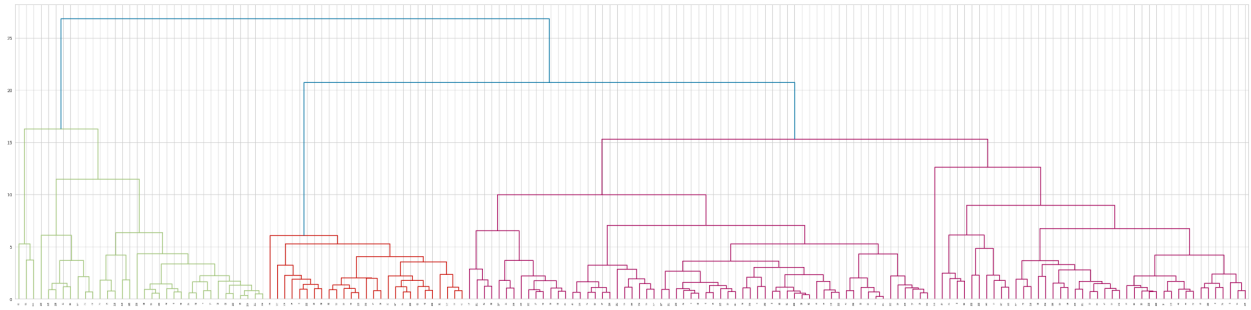
Then I trained the k-means clustering algorithm on PCA dataset with the number of clusters 3 .
Then we plotted the scatter plots of the clusters and divided the countries into 3 clusters i.e 0,1,2.
From the bar plots plotted we concluded that the countries in the cluster 0 are most underdeveloped.



We can observe that the K-means on the PCA dataset performs better than the normal dataset .

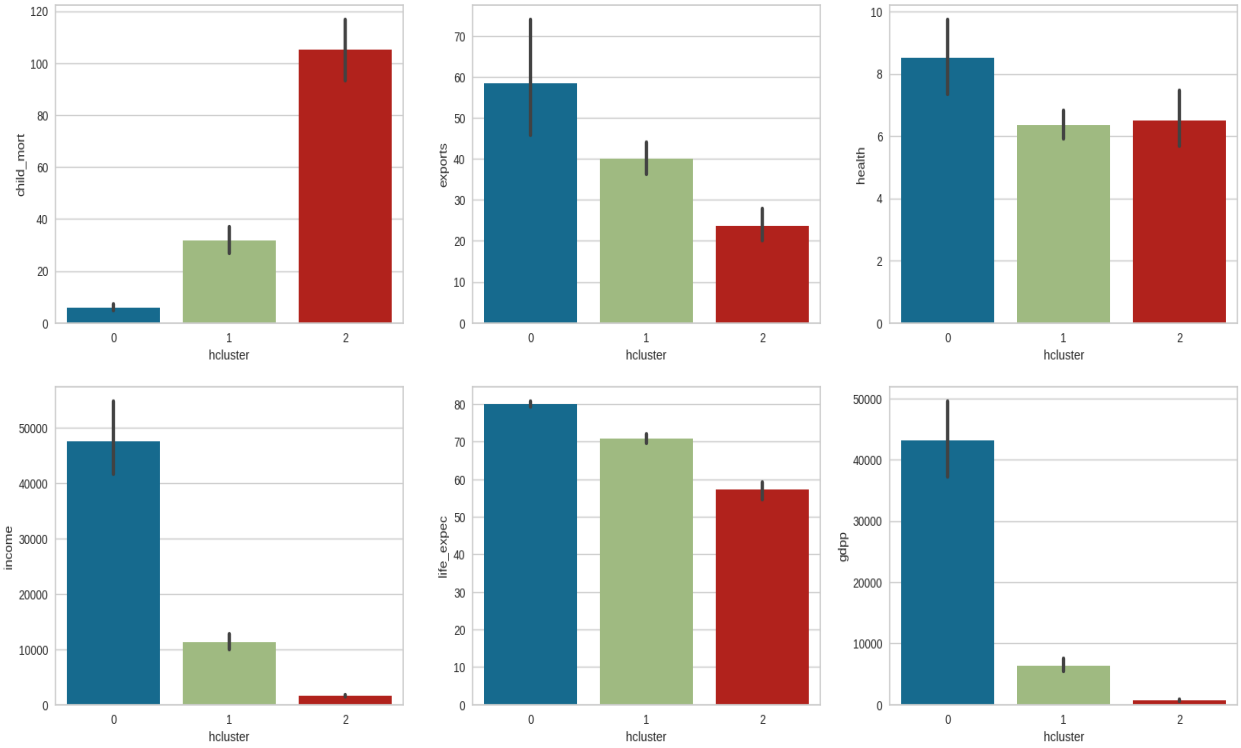
- **Hierarchical clustering:**

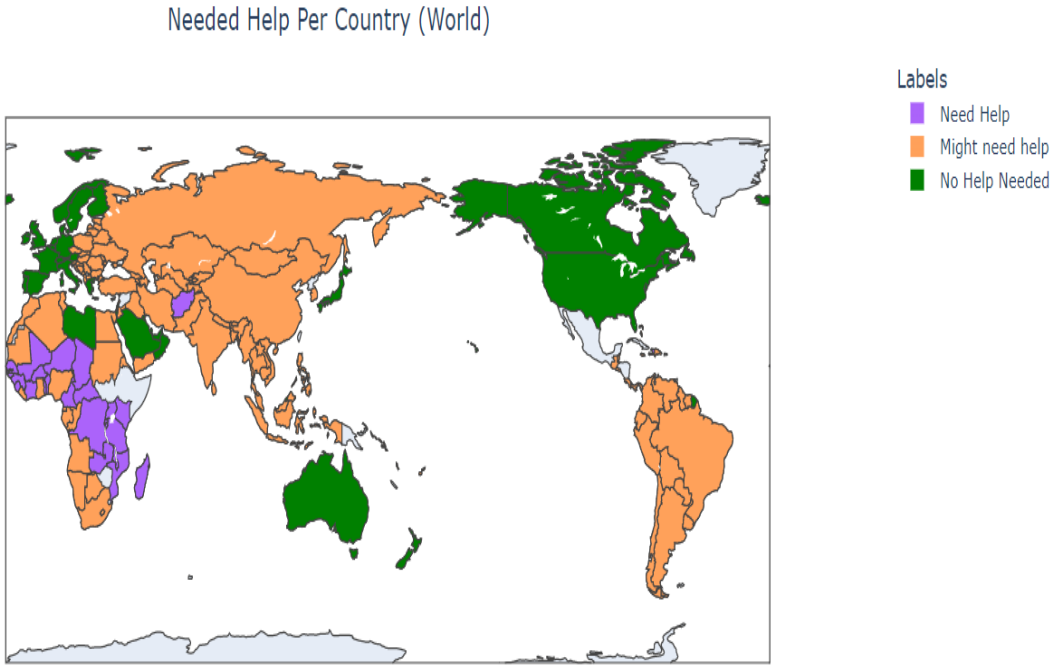
In this type of clustering , first we have made the dendrogram . By looking at the dendrogram we can say that the number of clusters is equal to 3.



Then we applied Agglomerative Clustering on the dataset.

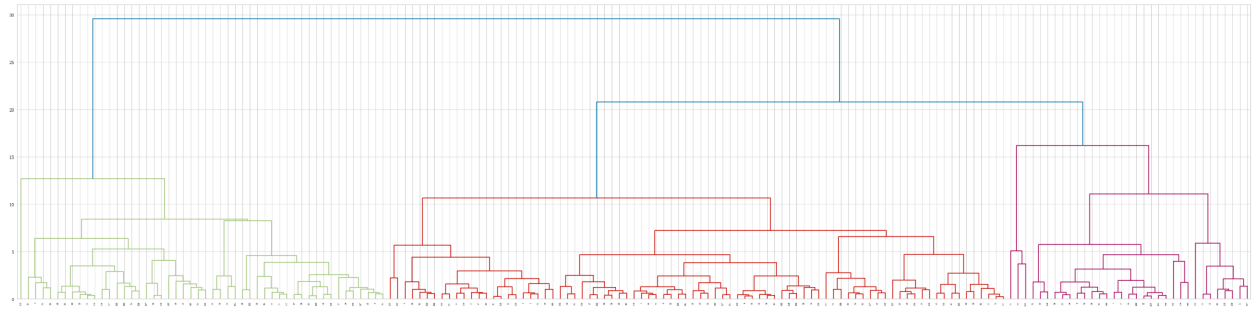
From the scatter plots and the bar plots plotted we concluded that the countries that are **clustered 2 are in need of the help** and we named them as **underdeveloped countries**, countries **clustered 1 are named as developing countries** and they might need help, but the **cluster 0** countries have performed poorly according to the criterion set above so they were named as **under-developed countries.**



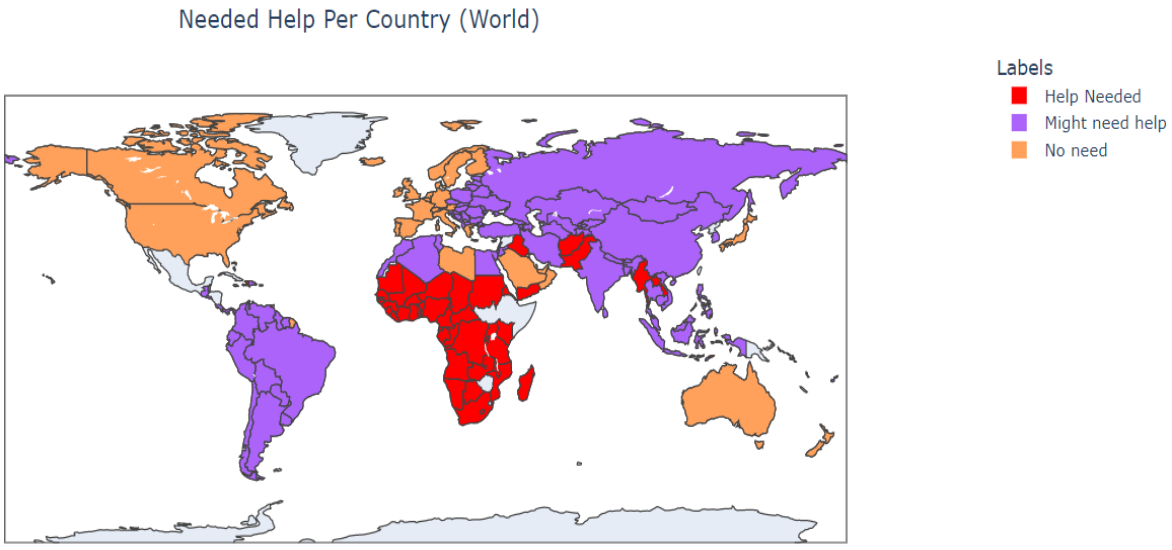


- **Hierarchical on PCA dataset:**

In this type of clustering , first we have made the dendrogram . By looking at the dendrogram we can say that the number of clusters is equal to 3.

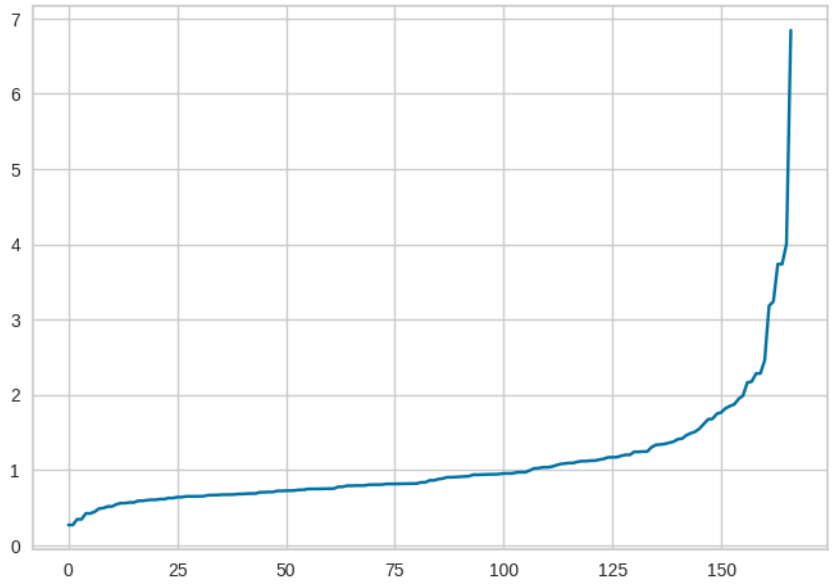


Then we applied Agglomerative Clustering on the dataset.
CLASS 0- DEVELOPED
CLASS-1 UNDER DEVELOPED
CLASS 2 DEVELOPING



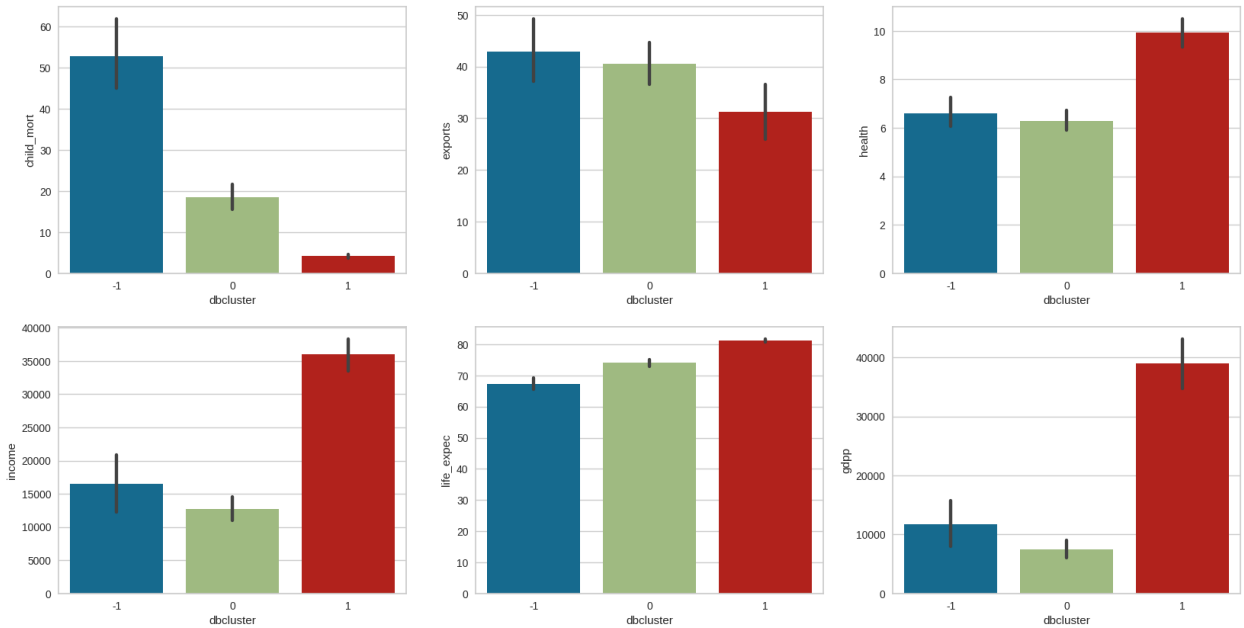
- **DBSCAN:**

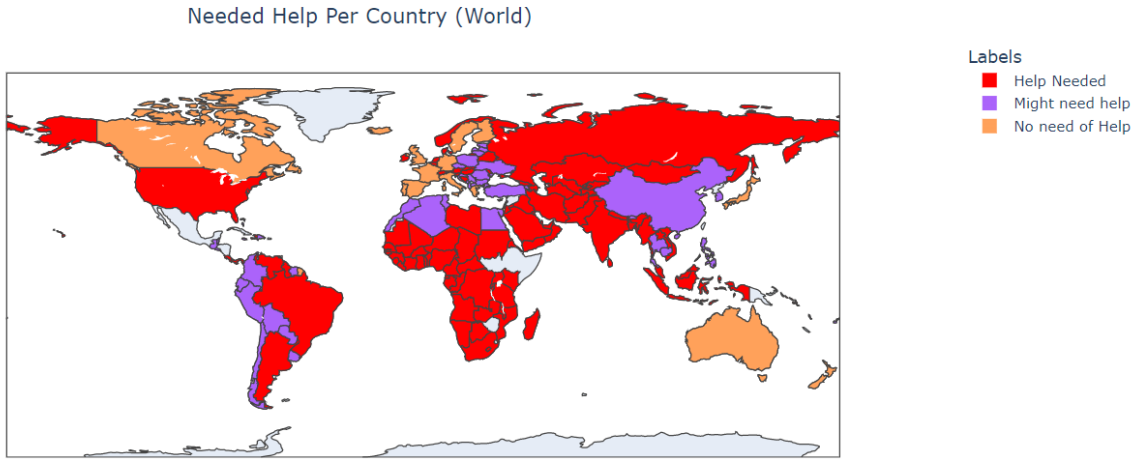
Now we applied the DBSCAN algorithm.
Firstly we did the hyperparameter tuning and selected the optimal epsilon value and min samples value.



Then we set the parameter values and applied the dbscan.
We observed that the number of cluster that were made by the algorithm was 3 in number and they were clustered according to the following:

CLASS 1- DEVELOPED
CLASS-1 - UNDER DEVELOPED
CLASS 0 - DEVELOPING





- **Conclusion:**

  In conclusion, we can say that hierarchical cluster performed the best among the three and DBSCAN performed the worst.
  The reason for Hierarchical clustering to perform this good is that Hierarchical clustering is a bottom-up approach that starts with individual data points as separate clusters and then merges them based on similarity, forming a tree-like structure called a dendrogram. This approach is useful when there is no prior knowledge about the number of clusters that exist in the data, as the dendrogram can be visually inspected to determine the optimal number of clusters. Additionally, hierarchical clustering can be more robust to noise and outliers since it considers all data points during the clustering process