CSL 2050
PATTERN RECOGNITION AND MACHINE LEARNING

Minor Project

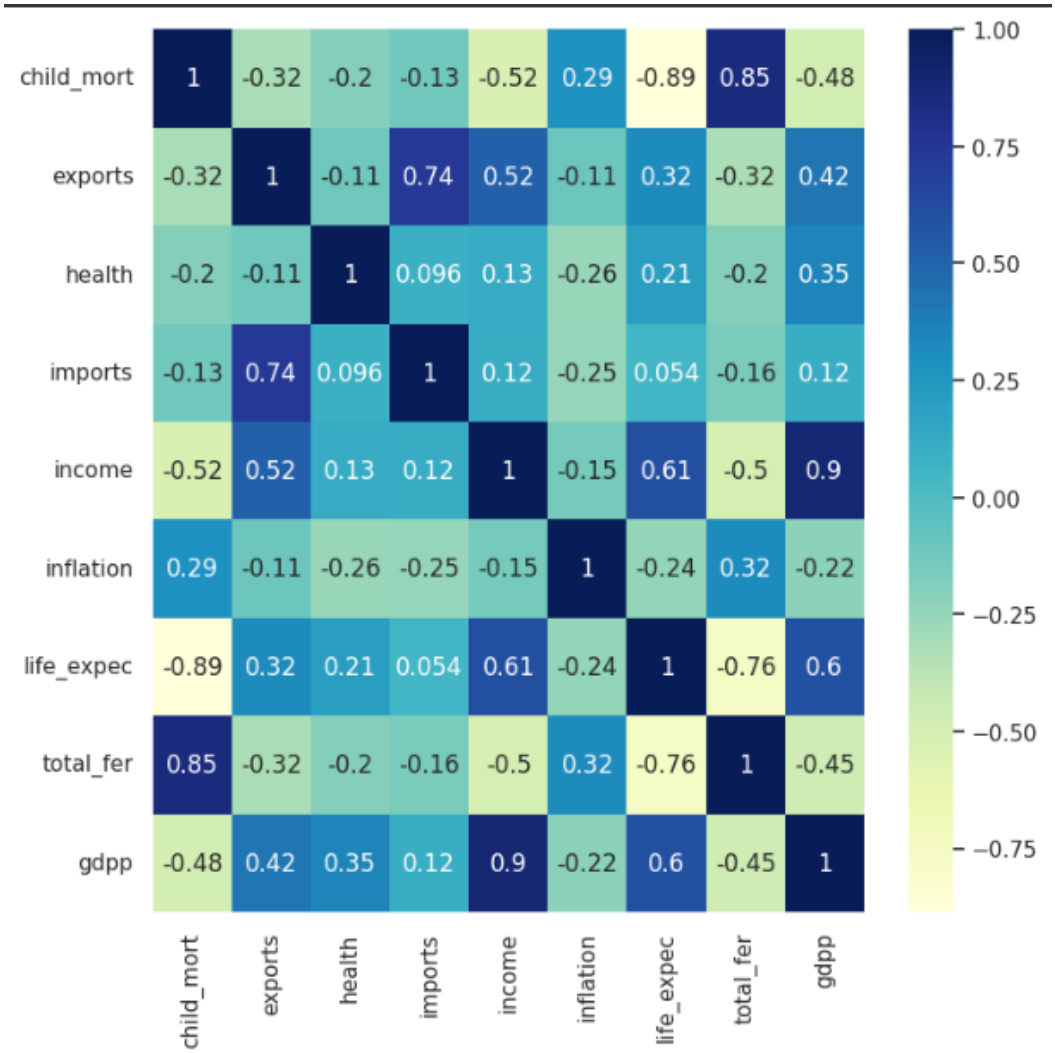
REPORT

Group Members :-
SHASHWAT ROY B21CS071
YOGESH JAHANGIR B21CS083
PRANAV PANT B21CS088

Dataset - Country DataSet

Part 1- Visualization

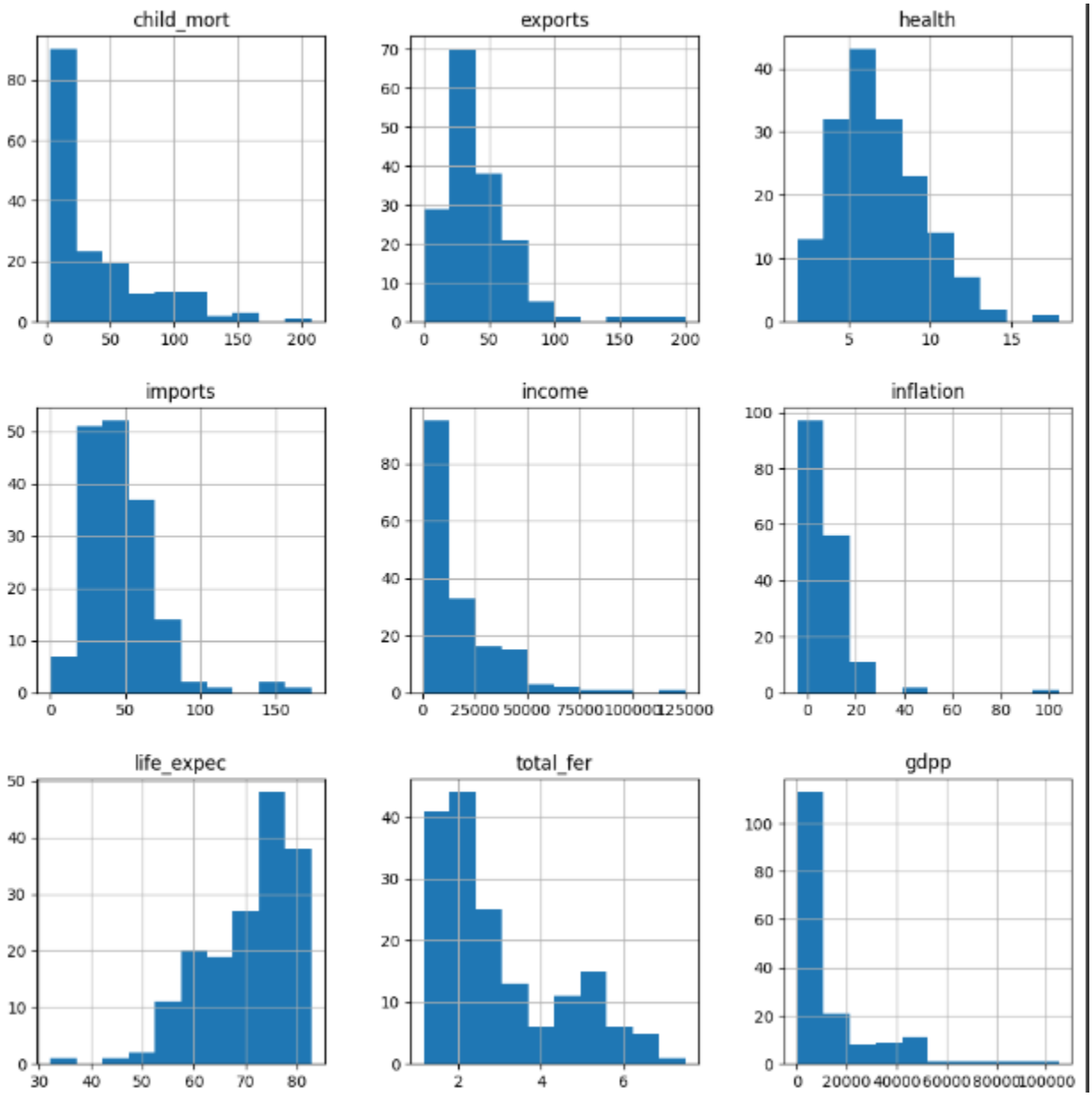
- Loaded the dataset as a dataframe and plotted some plots for visualization.
- Correlation Plot:-



- From the plot, we can see infer some things:
  - Child_mortality has a negative correlation with income, gdpp and a +ve correlation with total_fertility. We can say that when Income, gdpp increase, child mortality decreases. On the other hand, when total_fertility increases, child mortality increases(which is obvious).

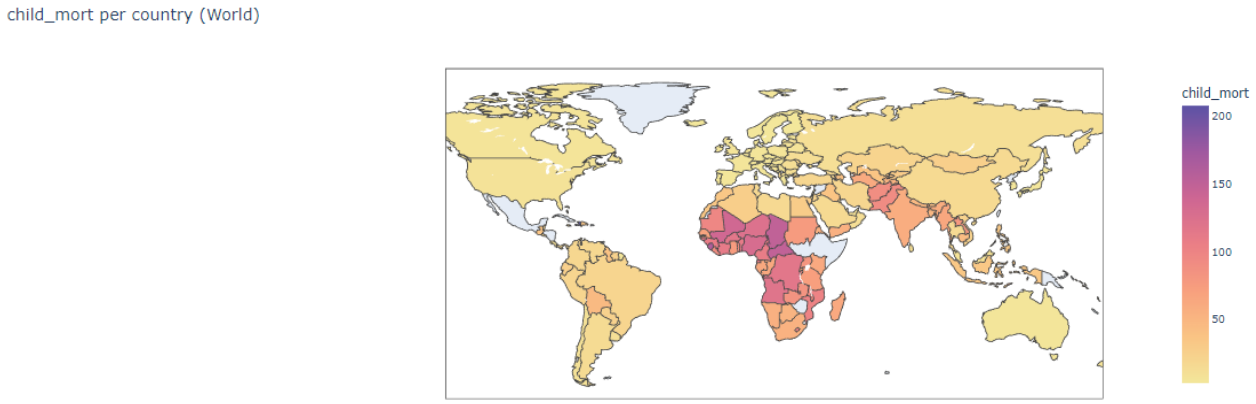
- Rise in income leads to rise in life-expectancy.
- Rise in exports increases gdp, income & imports.

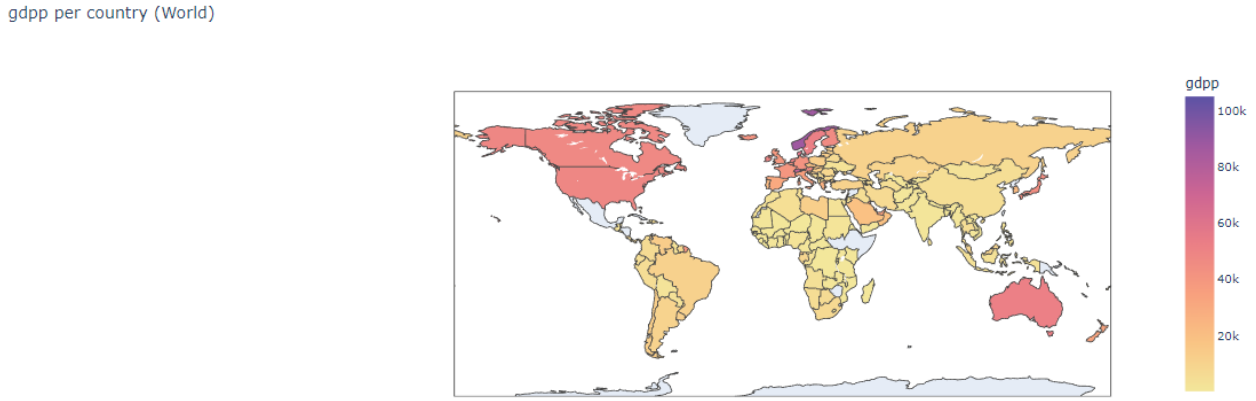
- Now we plot the data distribution using hist plots.



- From the plots, we can infer that many of the features are normally distributed, like exports, health, imports etc. and are +ve, -vely skewed, hence using standard scaler as scaling method would be best for this dataset.

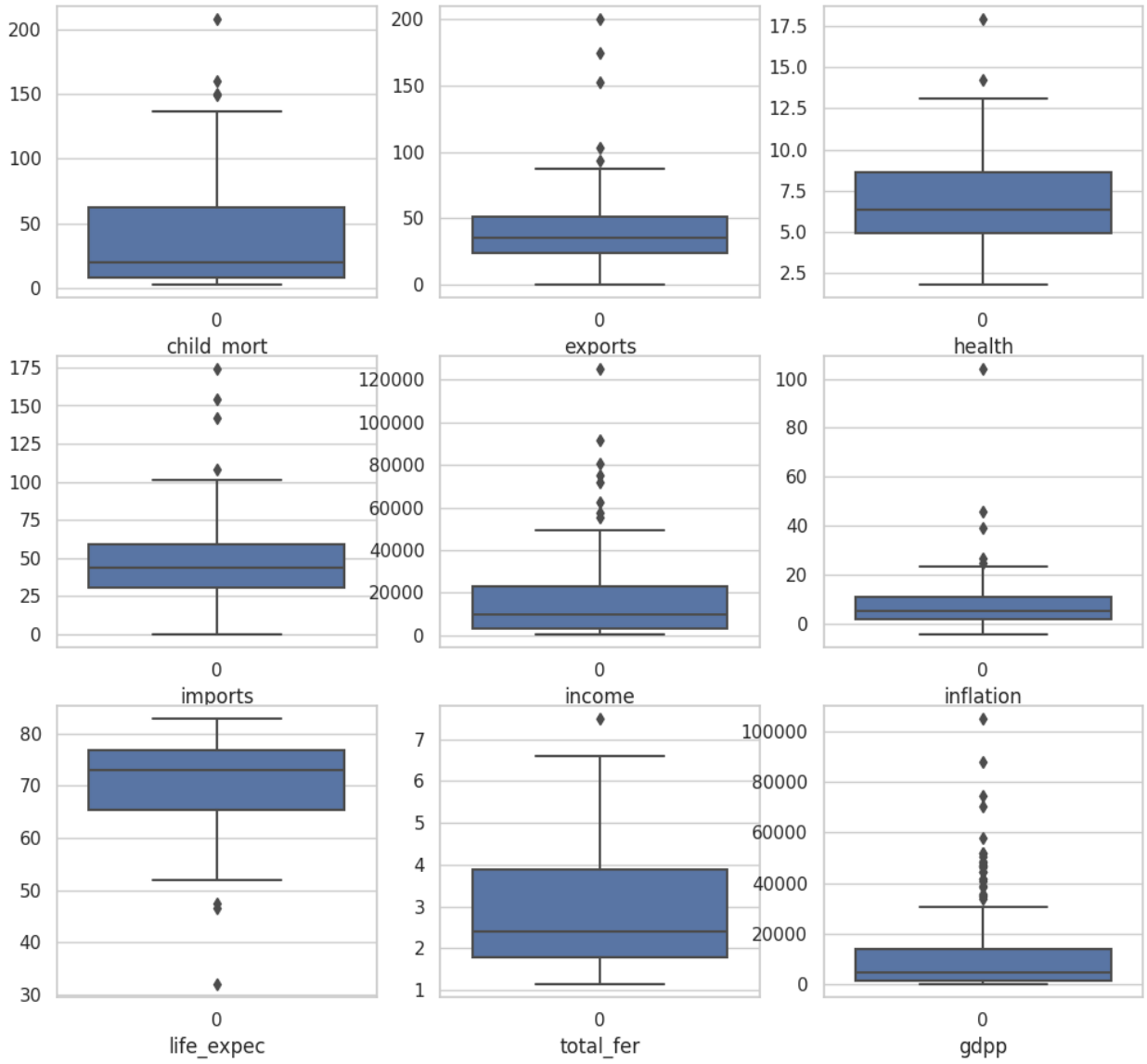
- Plotted the features on the world map using the *Plotly* module.





- Plotted similar plots for all the features (can be seen in the notebook file).
- Some inferences form these plots:-
    - Features like child mortality, inflation are high in poor countries, while income, exports and imports are low. These seem to be the characteristics of poor countries.
    - On the other hand, the countries with high GDP and Income spend good amount on health, have good exports and imports, and have a high life expectancy.

- Also plotted the boxplot for the features to look at the distribution.



Part 2- Preprocessing

- Checking for null values:

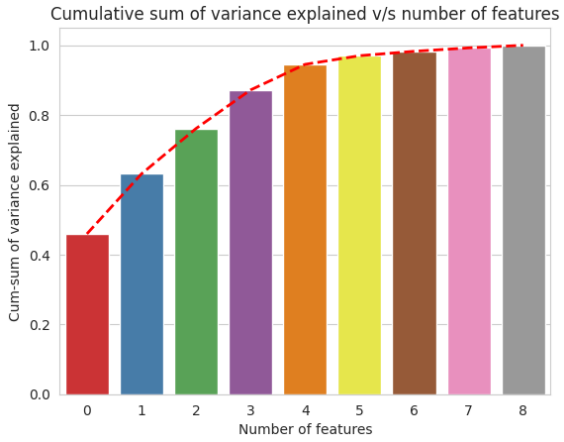
```
Number of null values in each column are:-

country       0
child_mort    0
exports       0
health        0
imports       0
income        0
inflation     0
life_expec    0
total_fer     0
gdpp          0
```

- Now, since all the countries are unique in the country column and there is no use of them if converted via scaling, hence we drop this column.
- Scaled the data using Standard Scaler.

Data Transformation techniques:

1. PCA Transformation :

   It is a linear dimension reduction technique which transforms the dataset to maximize the variance in that dataset.



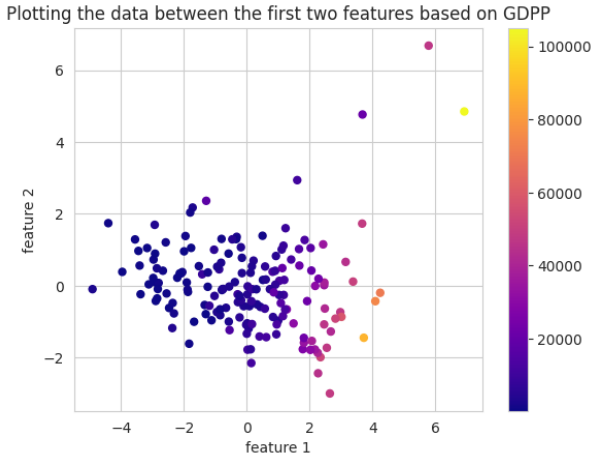
   The optimum value of n_components=6 for the PCA transformation.

2. t-SNE Transformation:

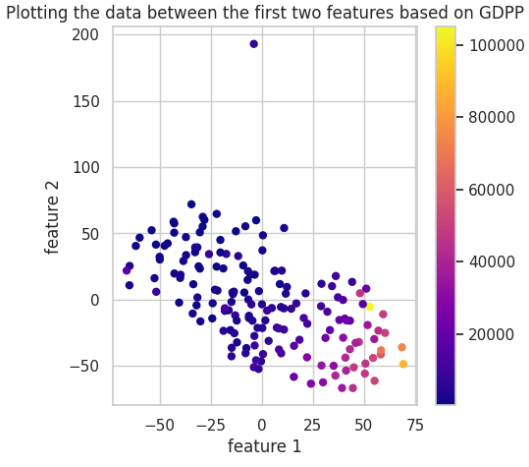
   It is a non-linear dimension reduction technique. Unlike PCA, it focuses on preserving the local variance in the distribution.

   Unlike PCA, it is hard to determine n_components for t-SNE. Moreover, it can only reduce the dimensions to below 4 components.

PCA transformed data



t-SNE transformed data



Part 3- **Model Training and Clustering**

For this, we considered 3 models:

1. Hierarchical Clustering

   Using dendrograms,hierarchical clustering allows for the identification of subclusters within larger clusters, providing a more detailed and nuanced understanding of the relationships between data points.

2. K-Means Clustering

   It is a simple clustering algorithm which assigns datapoints on the basis of shortest Euclidean distance to any cluster. In this dataset, we cannot treat any datapoint as ‘outlier’ so this clustering algorithm can be used.

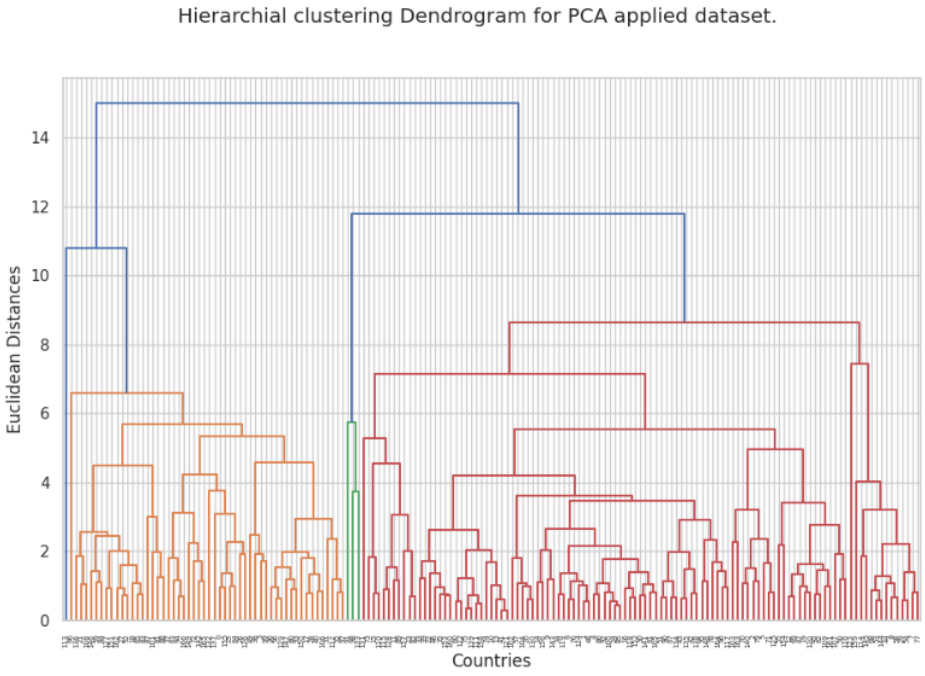
3. Fuzzy K-Means Clustering

   It is slightly different from K-Means Clustering in the sense that it is a softer clustering technique than K-means clustering and it allows a datapoint to belong to multiple clusters simultaneously and this is useful when a datapoint doesn’t perfectly align to a given cluster.

Hierarchical Clustering

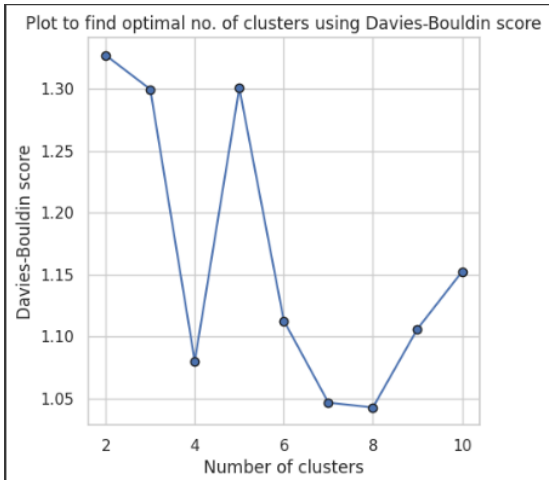
As we had transformed the dataset in two ways,t-SNE transformation and PCA transformation, we trained different K-Means models on both datasets.

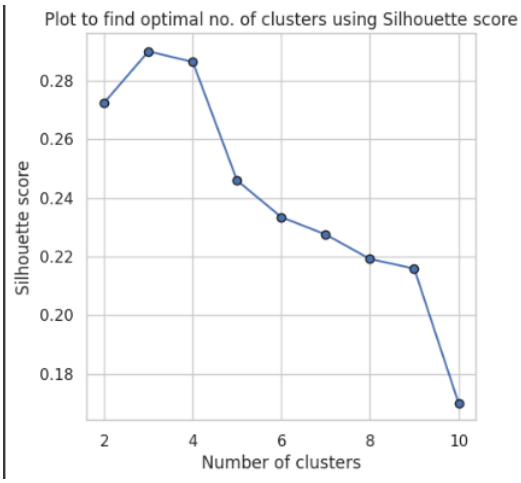
- First model trained on **PCA dataset**, we have the following plots for identifying the optimal number of clusters.



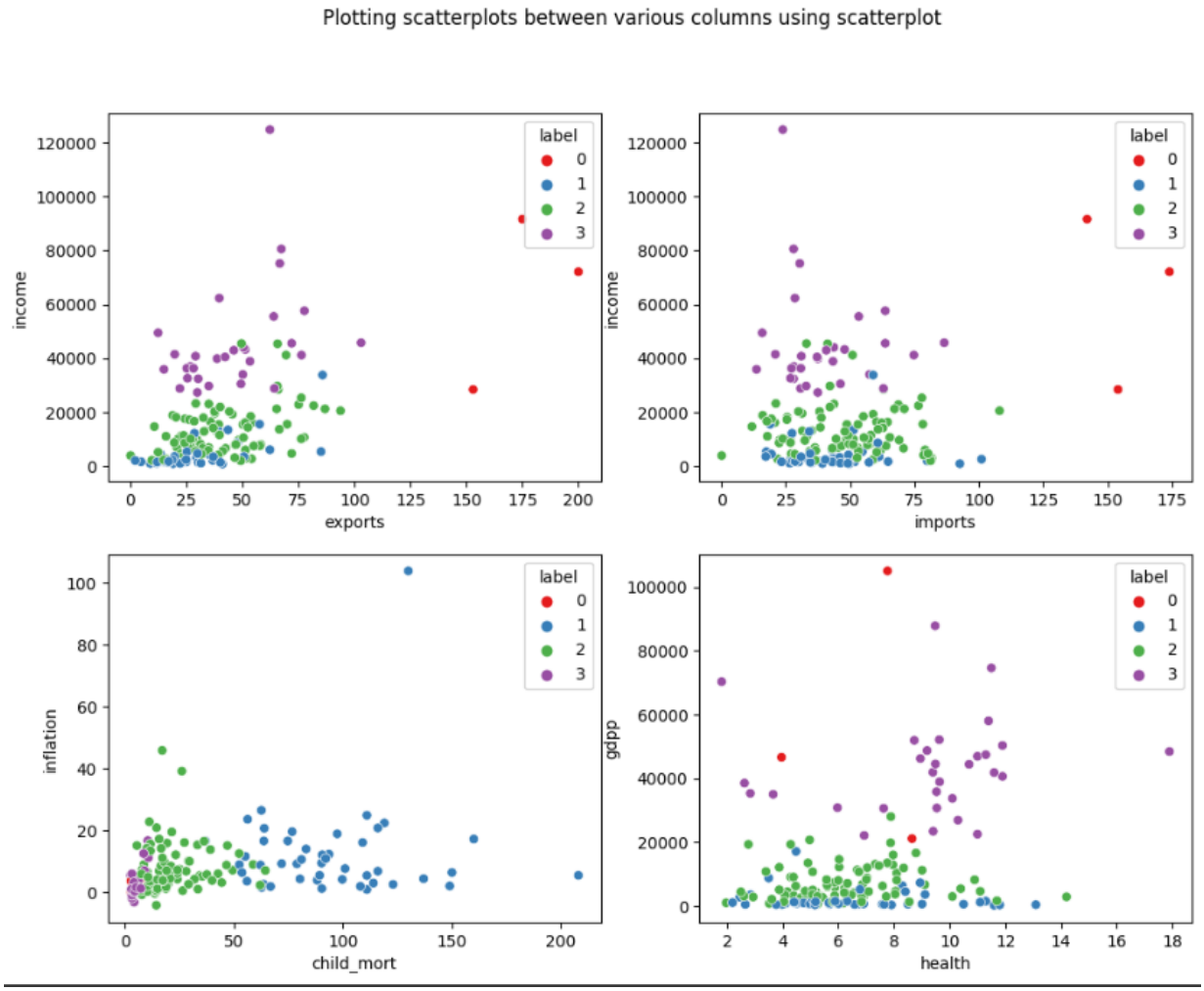
For getting the best number of clusters, we identify the longest vertical line that does not intersect any horizontal line. This line represents the greatest distance between any two points that are merged into a single cluster and the distance is called the "optimal distance."

We can see that the longest line is blue colored, and on cutting that, we get the optimal number of clusters =4.

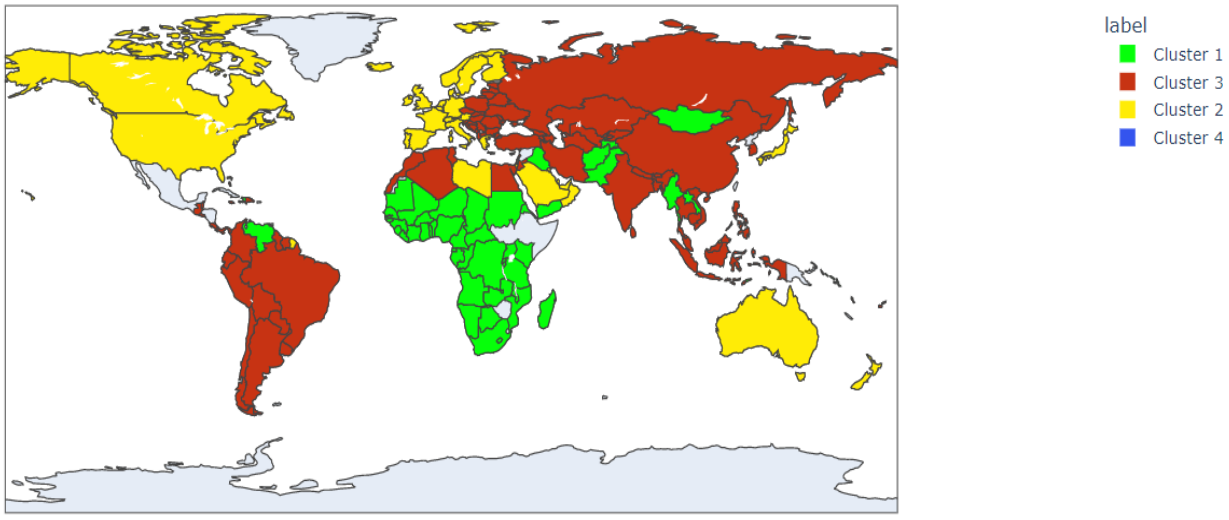




As a result, four clusters were created which can be visualized as follows:

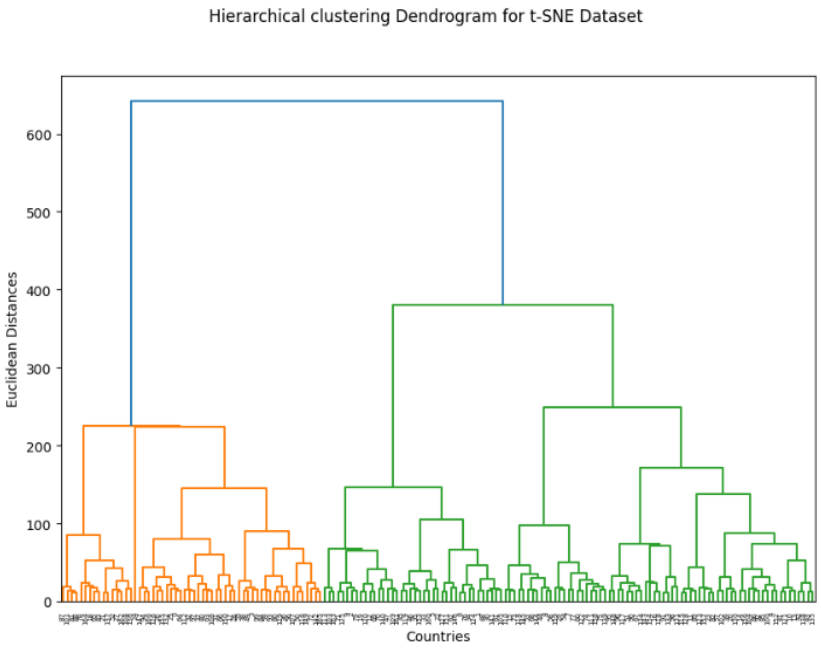


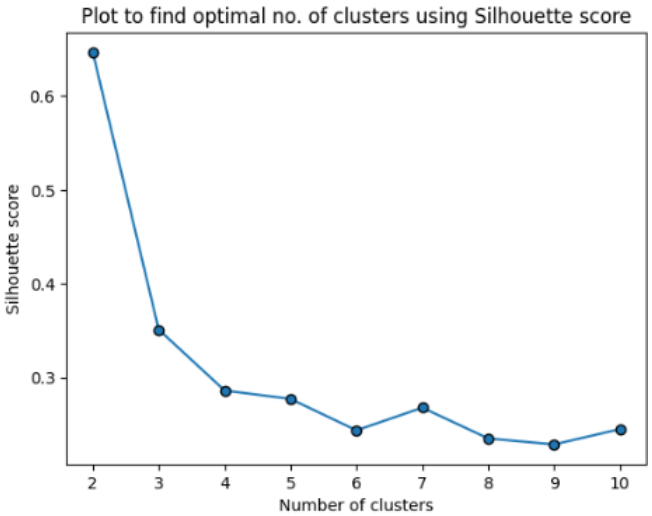
Countries clusters with Cluster 1 having the highest need for aid and Cluster 4 the lowest

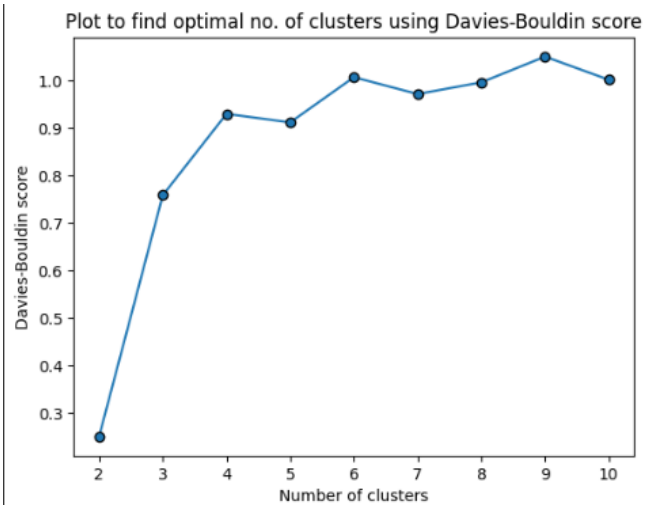


- Second model, trained on **t-SNE dataset**, we have the following plots**.**

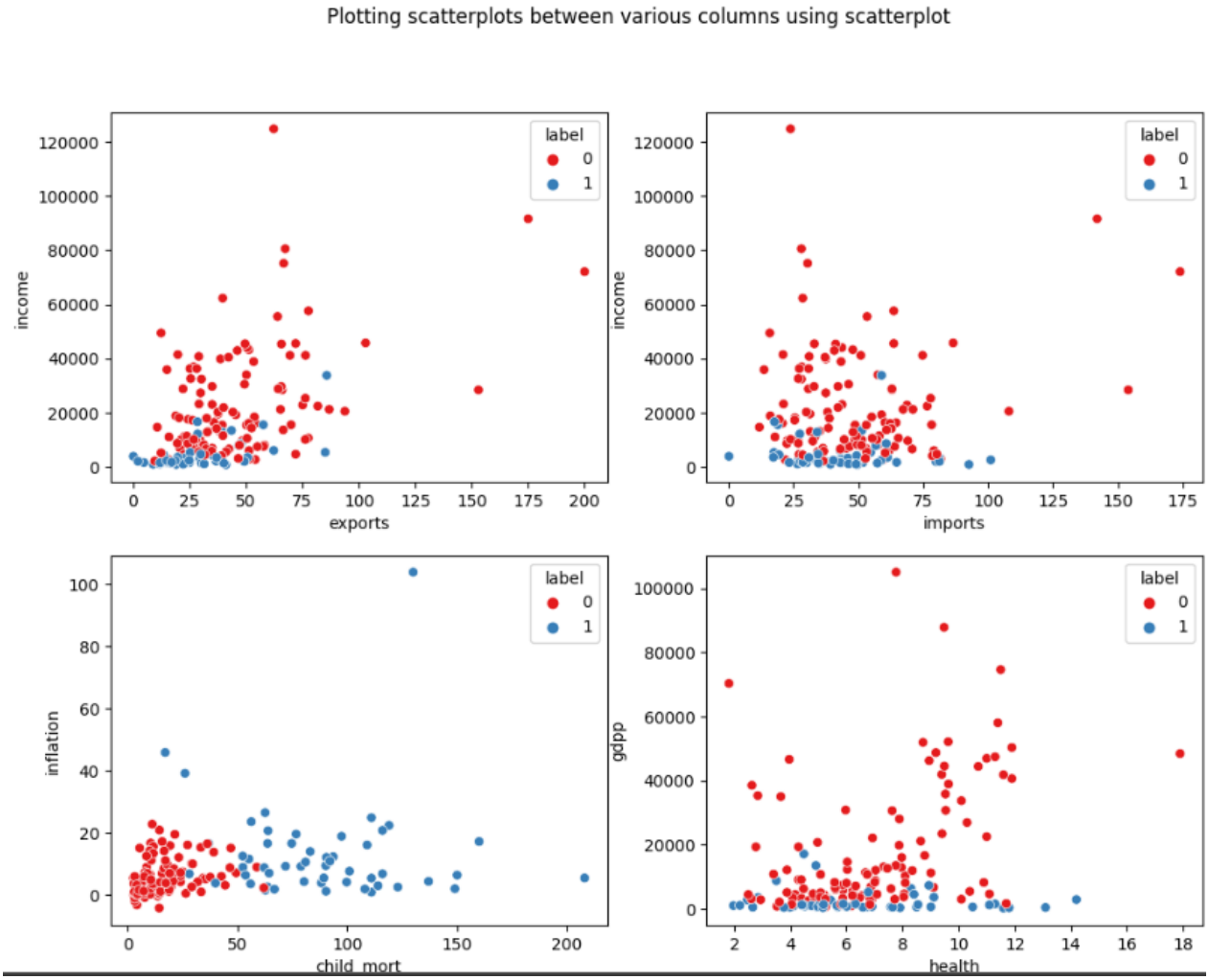
From the figure, optimal no. of clusters=2.



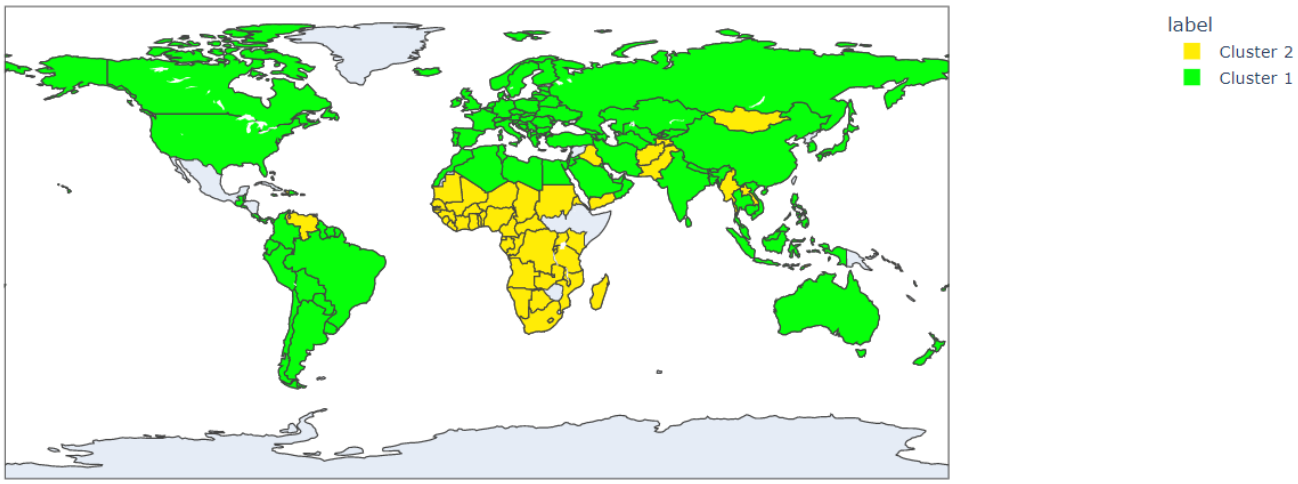




Using t-SNE transformation, we get optimal clusters=2. Plotting for n-clusters=2.



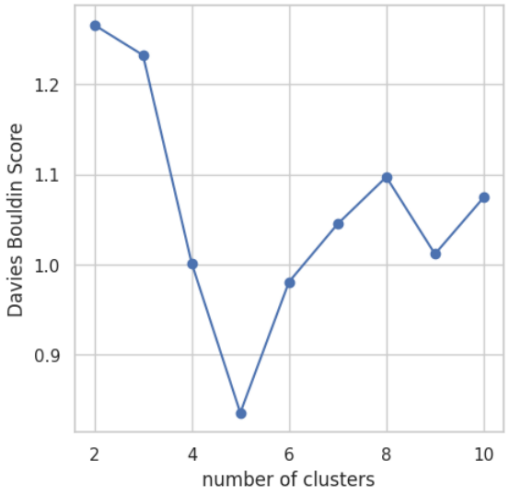
Countries clusters with Cluster 1 having the highest need for aid and Cluster 2 the lowest

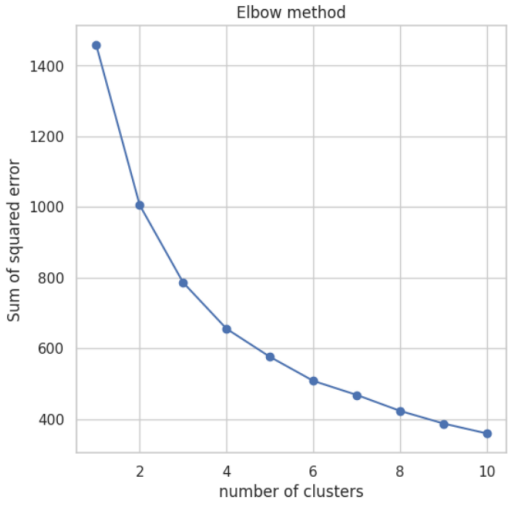


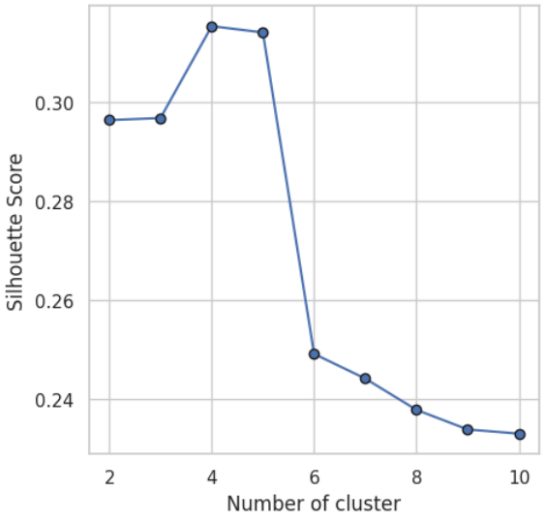
Since, there are only 2 optimal clusters for t-SNE, we can say that this is not optimal, as with 2 clusters, we are not able to fully differentiate between the countries. This is only a broad classification.
Major reason for obtaining 2 clusters could be that we applied t-SNE for n-components=3. It would not have been able to conserve all the variance, hence we got the clusters as 2.

K-Means Clustering

Since we had transformed the dataset in two ways,t-SNE transformation and PCA transformation, we trained different K-Means models on both datasets.
The first model was trained on the PCA dataset.

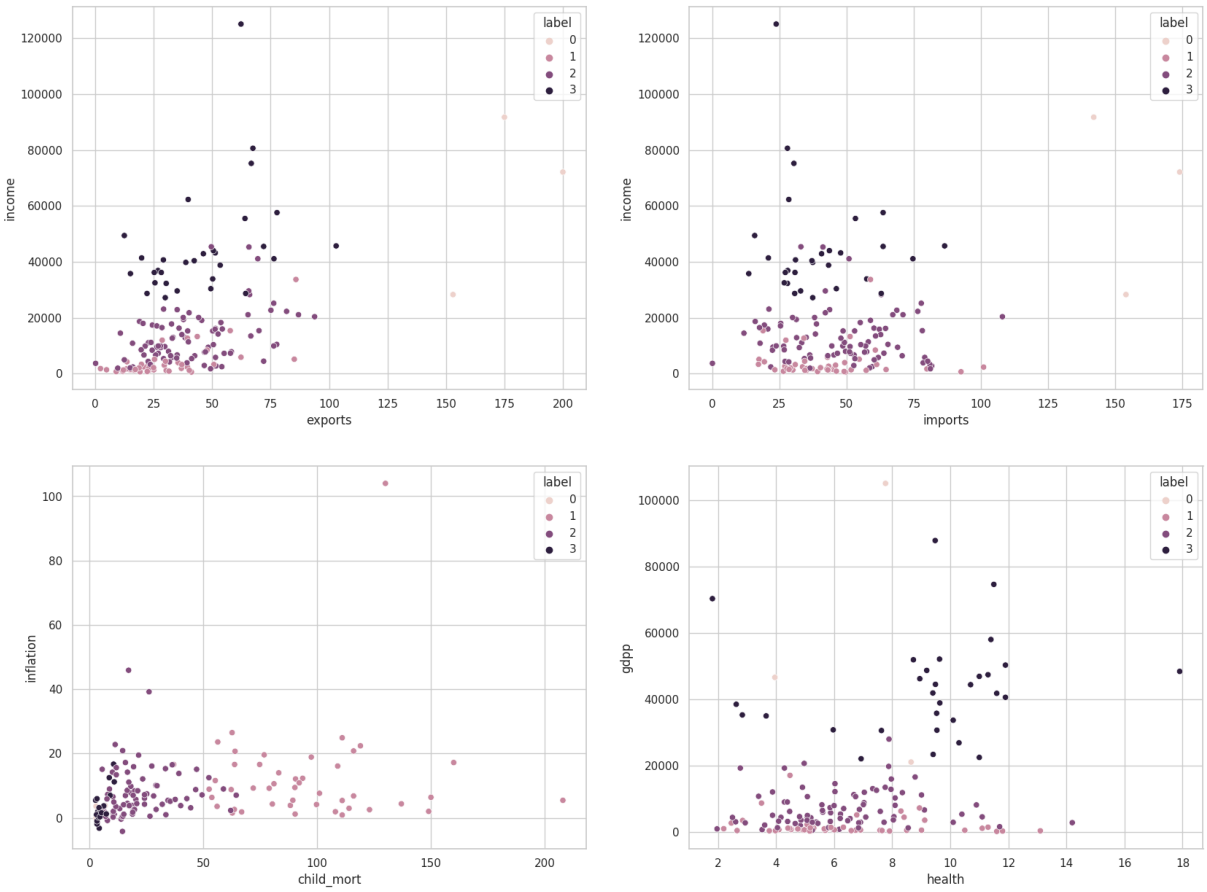




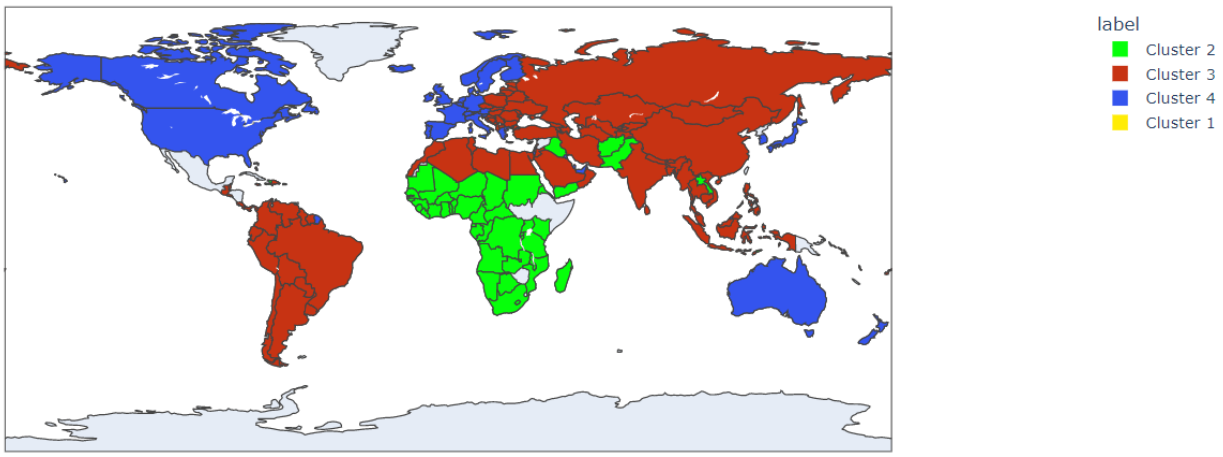


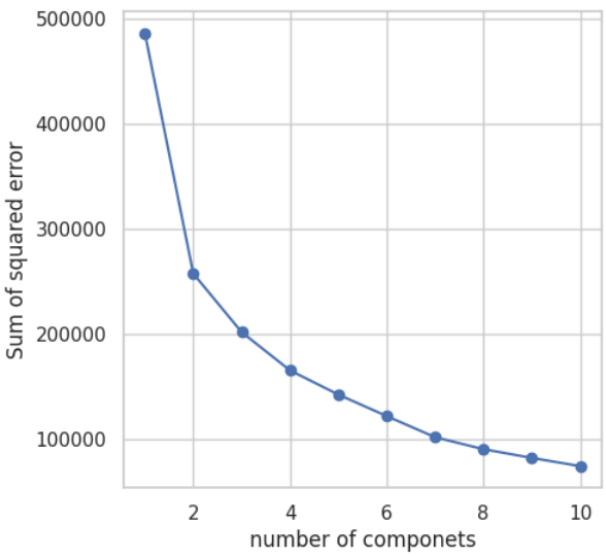
Using the methods shown above, we adjudged the value n_clusters=4.
As a result, four clusters were created which can be visualized as follows:

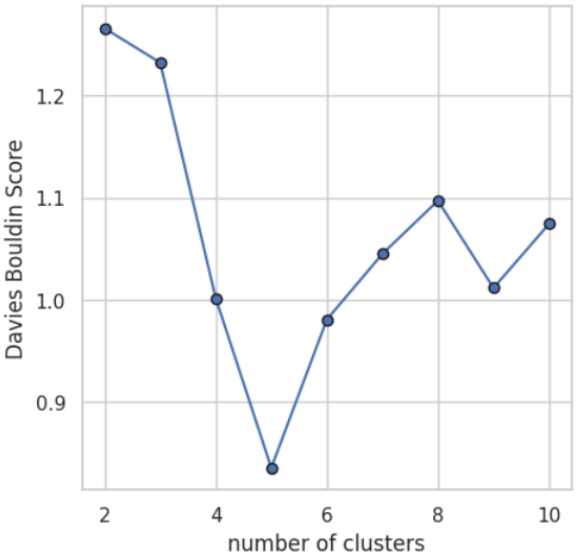
Plotting scatterplots between various columns using scatterplot



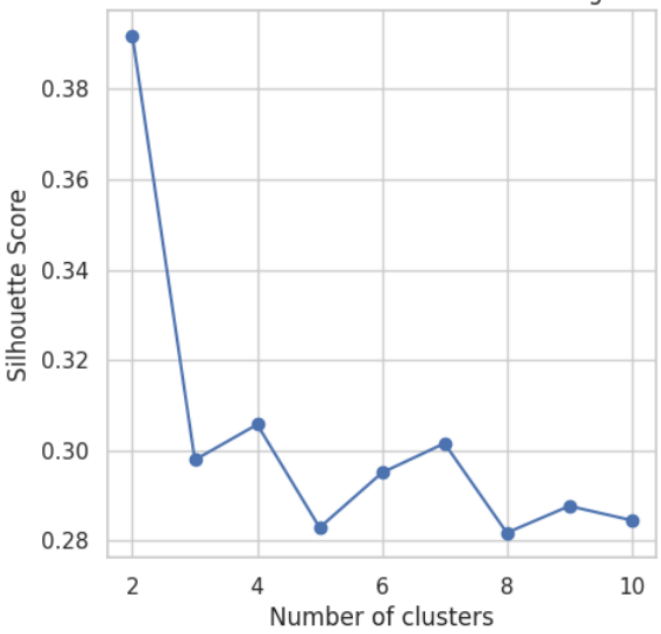
Countries clusters with need for aid rising in the order: Cluster 1,Cluster 4,Cluster 3,Cluster 2





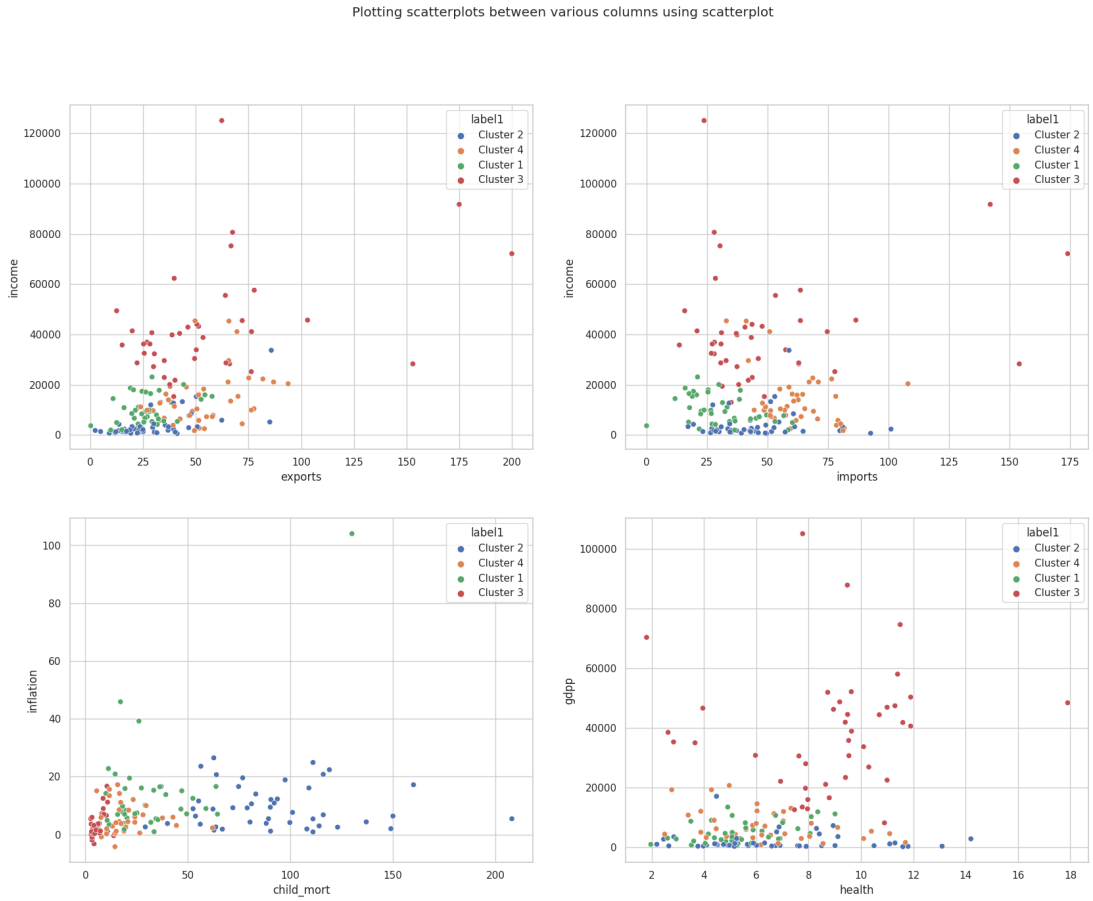


Plot to find best number of clusters from k-means clustering using Silhouette score

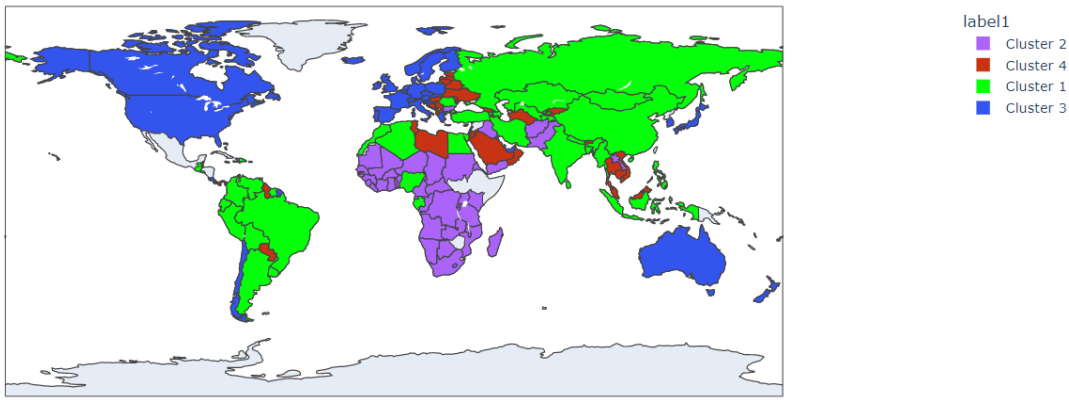


Using the methods shown above, we adjudged the value n_clusters=4.

As a result, four clusters were created which can be visualized as follows:



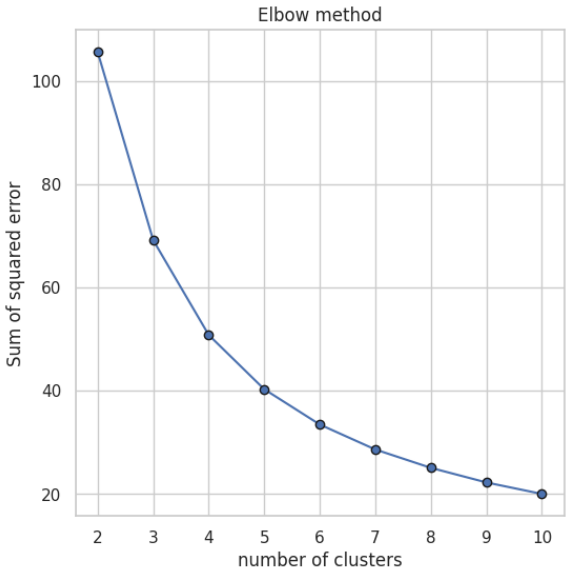
Countries clusters with need for aid rising in the order: Cluster 3,Cluster 4,Cluster 1,Cluster 2

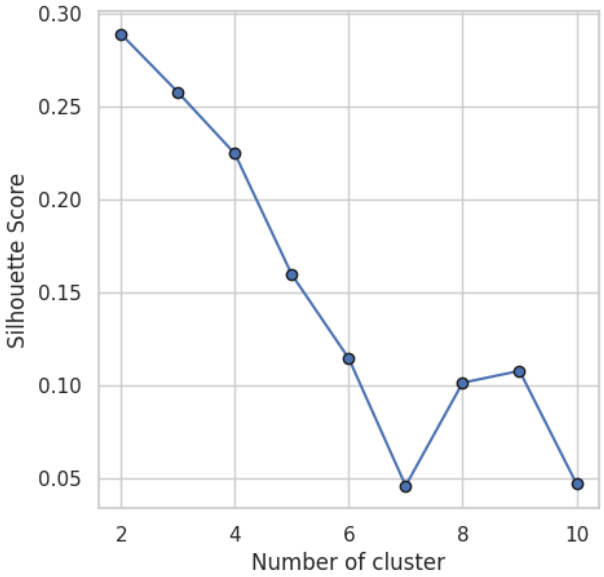


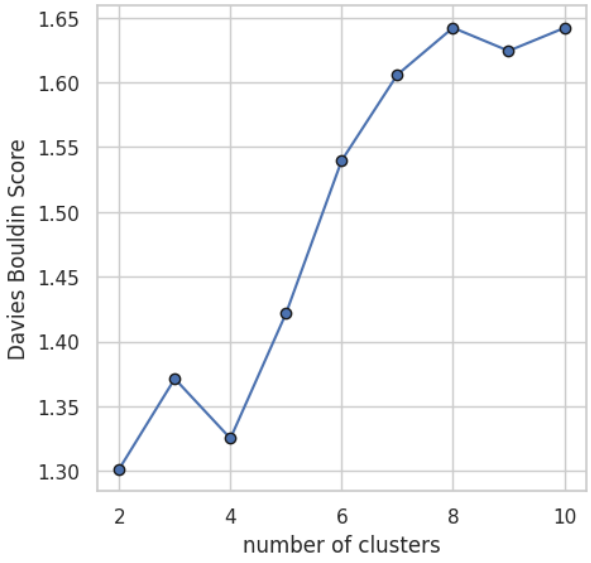
Fuzzy K-Means Clustering

Since we had transformed the dataset in two ways,t-SNE transformation and PCA transformation, we trained different K-Means models on both datasets.

The first model was trained on the PCA dataset.

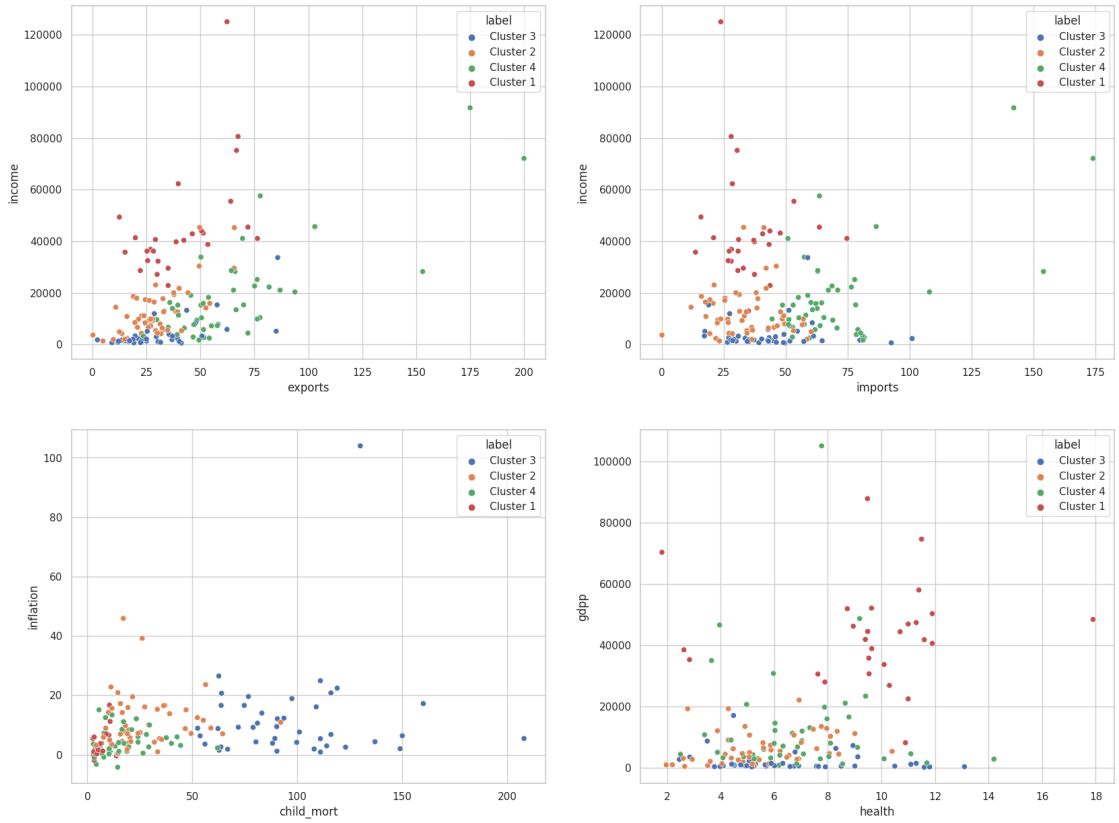




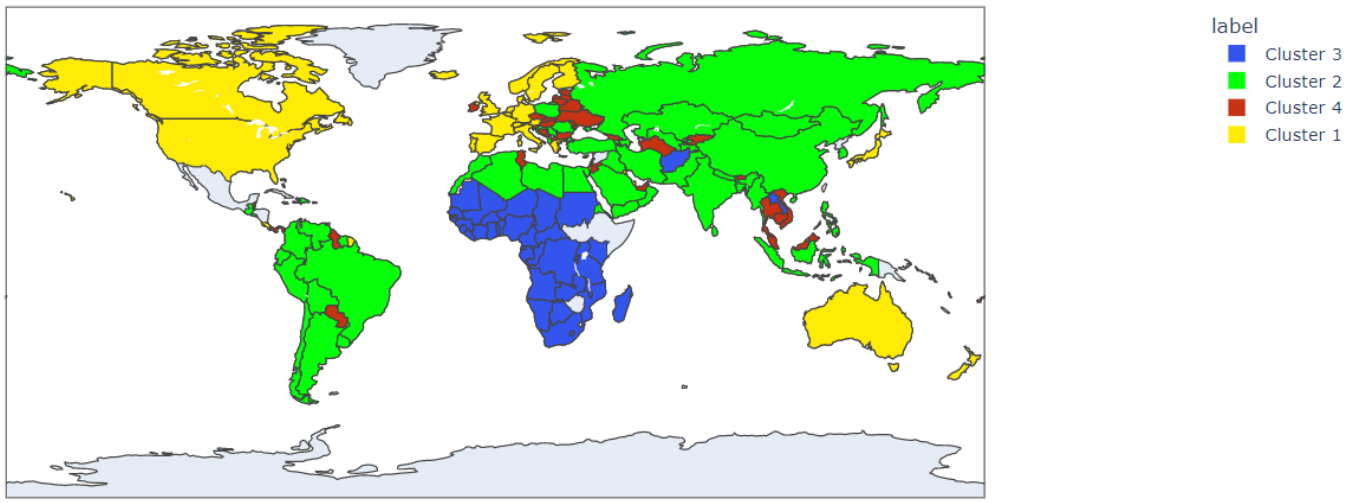


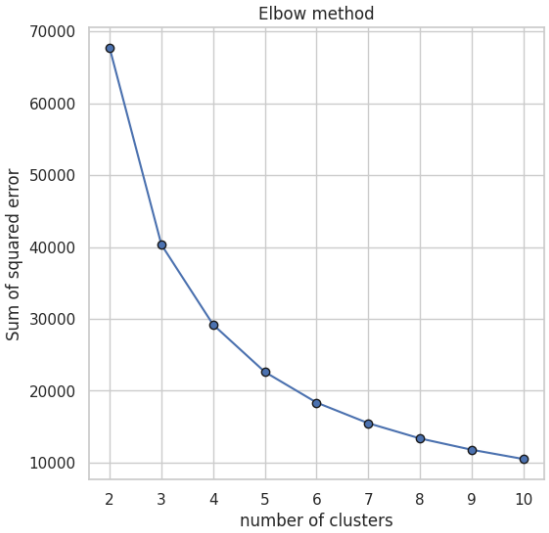
Using the methods shown above, we adjudged the value n_clusters=4.
As a result, four clusters were created which can be visualized as follows:

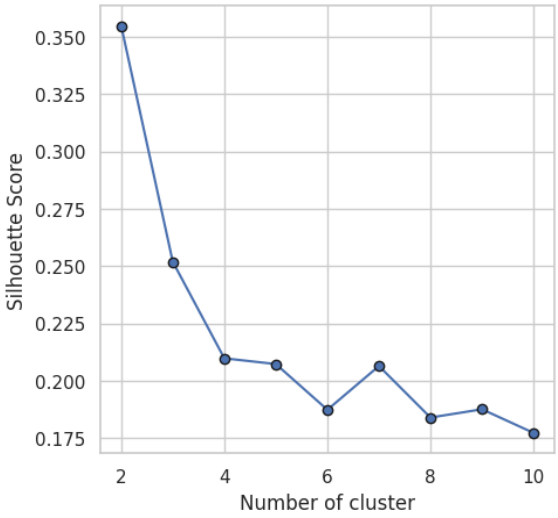
Plotting scatterplots between various columns using scatterplot

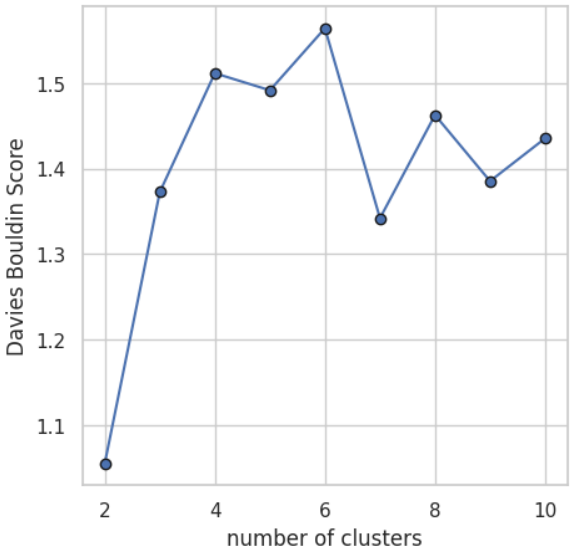


Countries clusters with the need for aid rising in the order : Cluster 1,Cluster 4,Cluster 2,Cluster 3



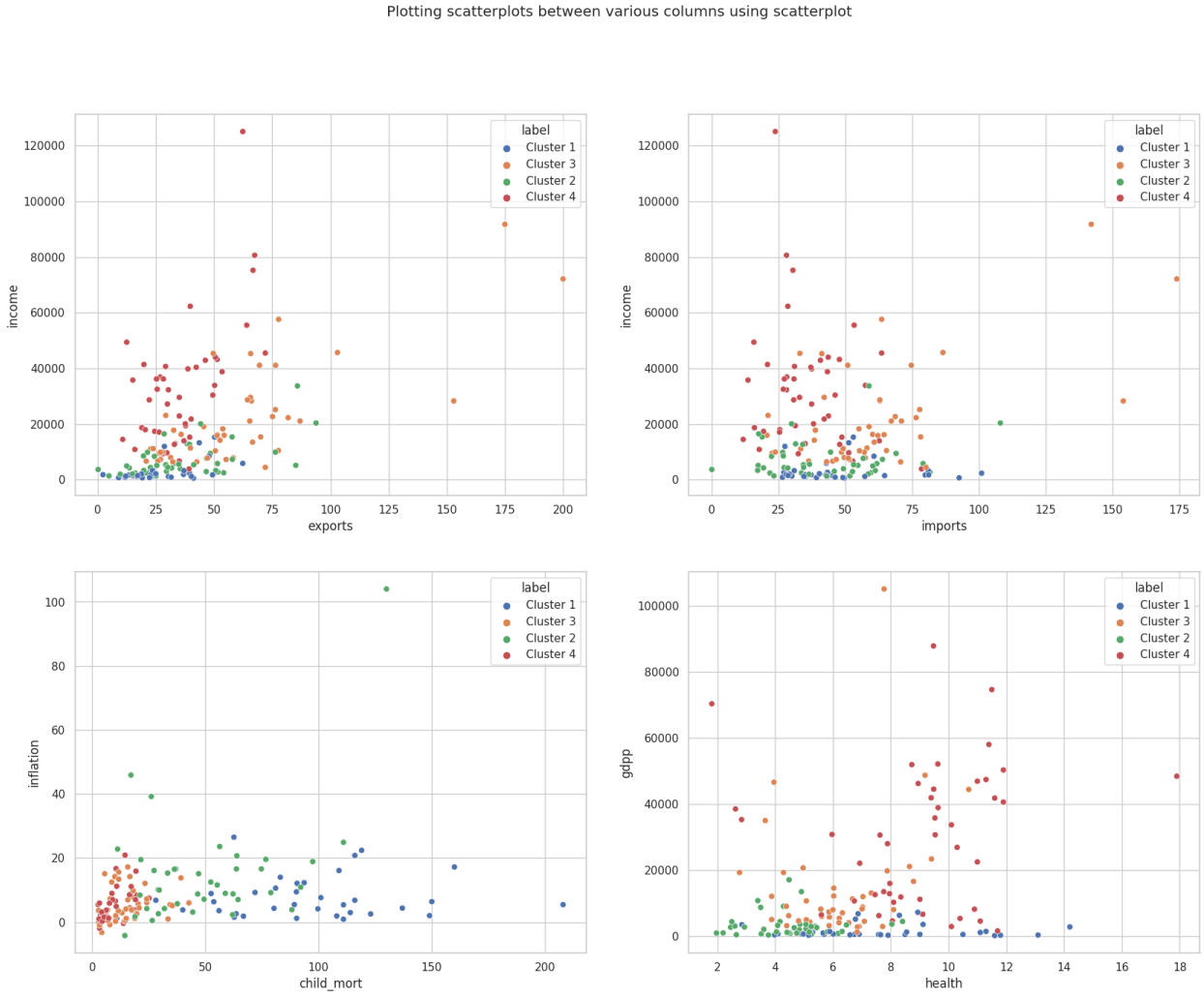




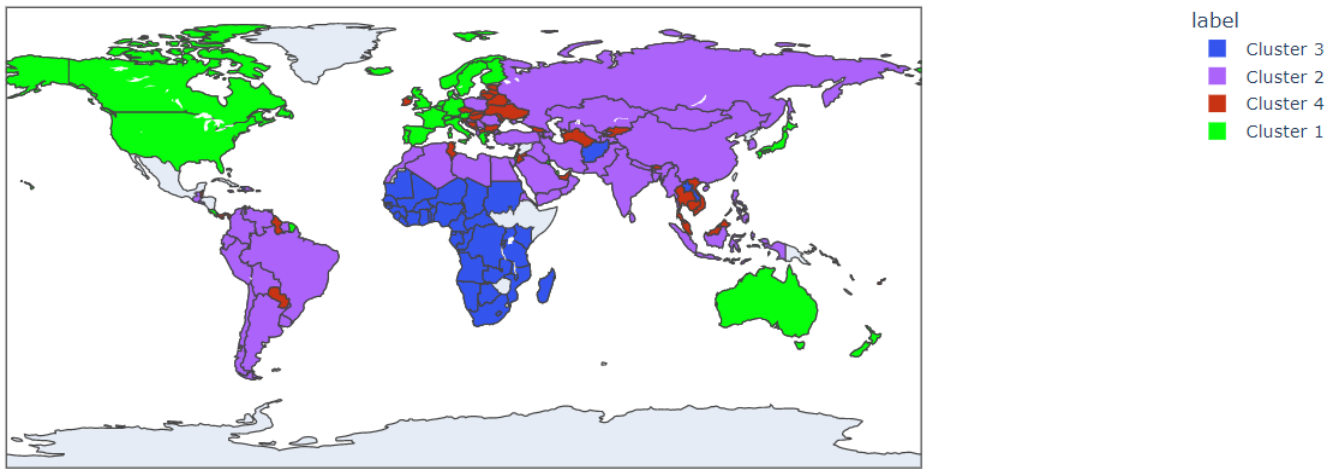


Using the methods shown above, we adjudged the value n_clusters=4.

As a result, four clusters were created which can be visualized as follows:



Countries clusters with the need for aid rising int the order: Cluster 4,Cluster 3,Cluster 2,Cluster 1



Part 4 - Observations and Conclusions

Parameters	Hierarchical Clustering		K-Means Clustering		Fuzzy K-Means Clustering	
	PCA	t-SNE	PCA	t-SNE	PCA	t-SNE
Silhouette Score	0.308	0.646	0.315	0.306	0.224	0.210
Davies-Bouldin Score	1.046	0.250	1.002	1.002	1.325	1.512
No. of clusters	4	2	4	4	4	4

From the observations on the scores listed above, we can infer that Hierarchical Clustering on t-SNE transformed dataset gives the best clusters. However, since there are only 2 clusters formed, it only gives us a broad and ambiguous categorization of the countries which is not suitable.

The next best clustering was found for the K-Means Clustering algorithm on the PCA transformed dataset. With 4 clusters being produced, it gave us a much clearer categorization of the countries.