॥ त्वं ज्ञानमयो विज्ञानमयोऽसि ॥

PROJECT REPORT

Unsupervised Learning of Country Data Using Machine Learning

Team Members -

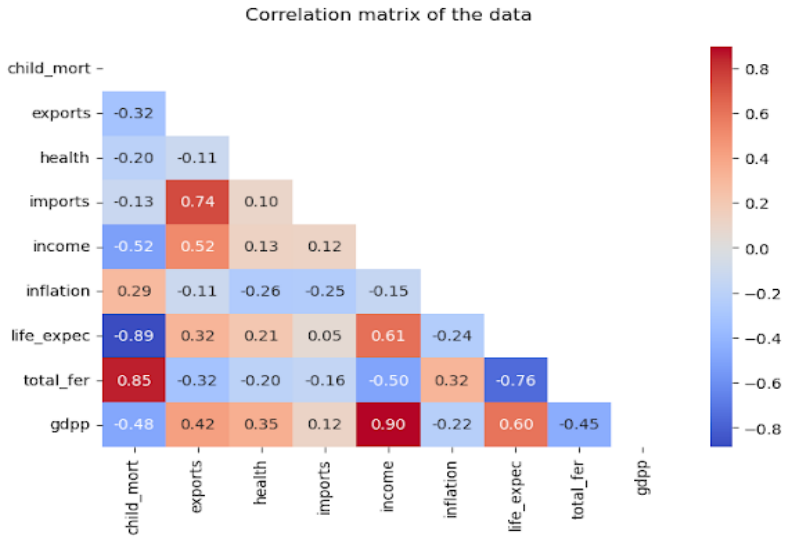
- Anushkaa Ambuj (B21ES006)
- Jyoti Dhayal(B21EE029)
- Anupama Birman(B21EE008)

Data Pre-Processing

- Checking NULL values
- Feature Scaling: We have implemented three features scaling techniques
    - Standard Scaling – scaled all the data using standard scaler
    - Min max scaling – scaled all the features using Min Max scaling
    - Combination of standard scaling and Min Max scaling – scaled 'Health' feature using standard scaler as it was normally distributed and other features using Min Max scaler.

Data Visualization

Correlation Matrix

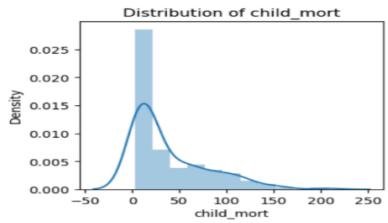


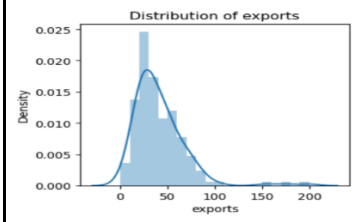
Observations

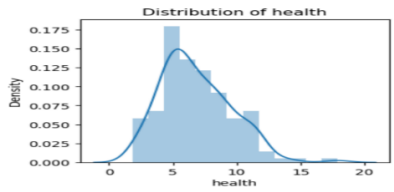
According to above heatmap:

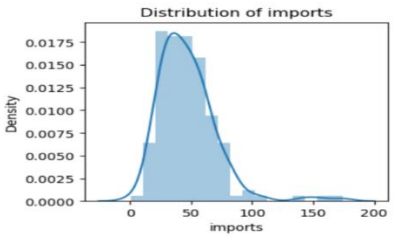
- Income ↑, gdpp ↑ ----> Child-mort ↓
- Child-mort ↑ ----> Life-exp ↓
- Total-fer ↑ ----> Child-mort ↑
- Export ↑ ----> Income ↑
- Income ↑ ----> Life-exp ↑
- Income ↑ ----> Total-fer ↓
- Life-exp ↑ ----> Total-fer ↓
- Life-exp ↑ ----> GDPP ↑

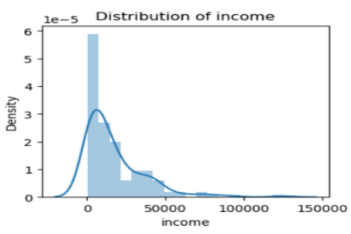
Feature Distribution

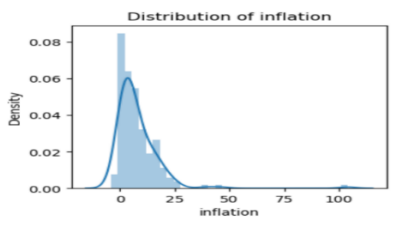


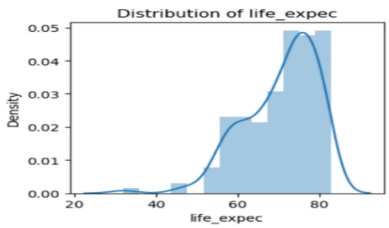


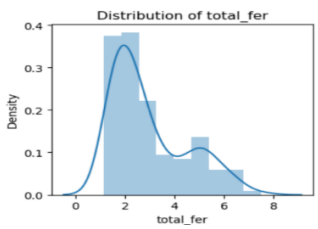


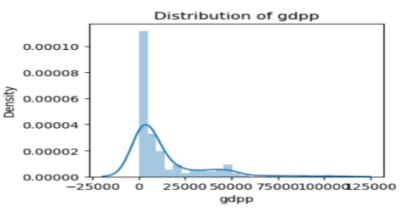










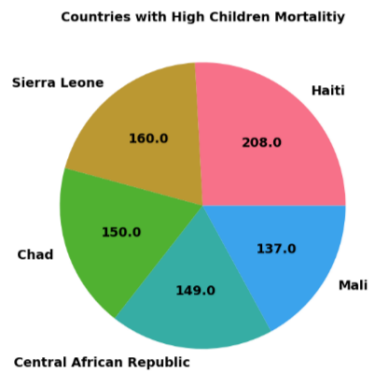


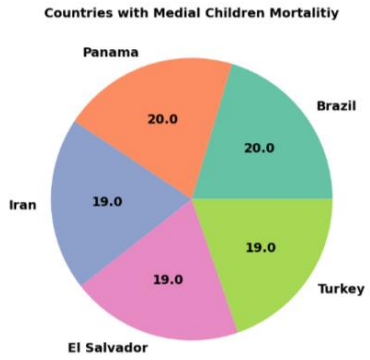
Observations

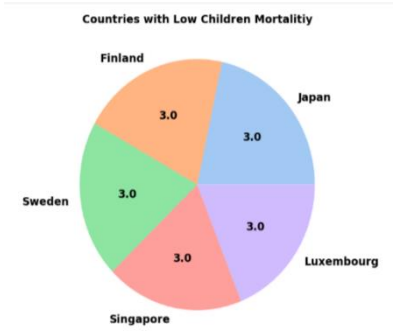
- *Health data follows a normal distribution.*
- *Life-exp is skewed to the left.*
- *The rest are skewed to the right. (positive skewness)*

Pie Charts

a) Child Mortality

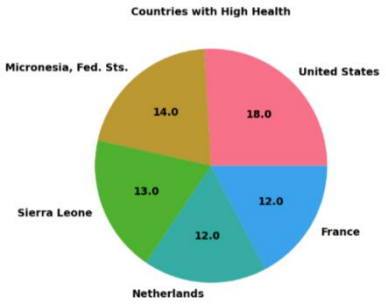


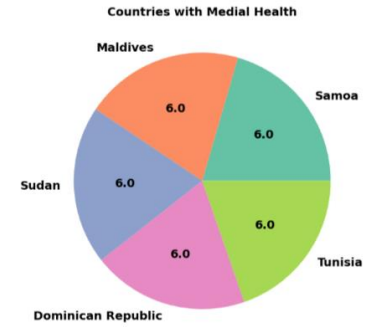


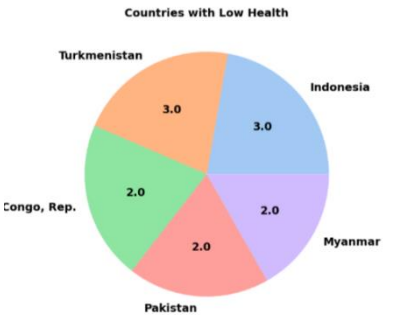


- *Haiti, located in Central America, has the highest number of child mortality. Most African countries are also seen in these statistics.*
- *On the other hand, European countries and some Asian countries have the lowest child mortality rate*

b) Health



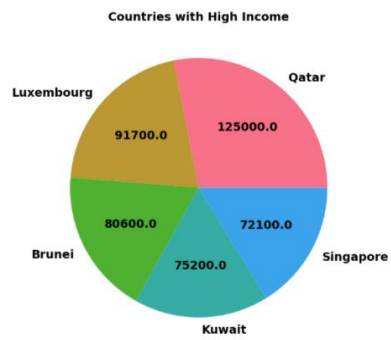


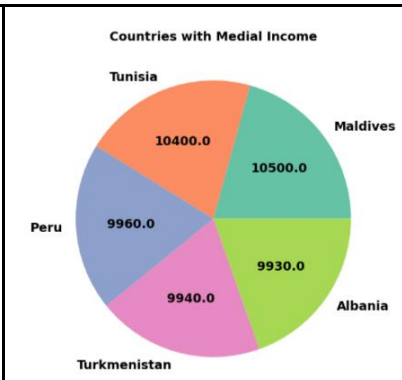


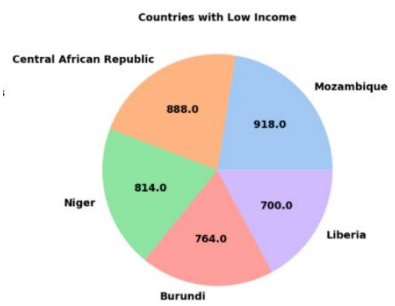
- *The USA is the country that spends the most on health.*
- *Countries that pay less attention to health issues are located in the Asian continent.*

- *Singapore, Malta, Luxembourg, and Seychelles are on the list of countries that import the most. These countries also had the largest exports.¶*
- *Brazil is the least importer among the countries. Sudan, Argentina, USA, and Japan are among the countries that have low imports.*

c) Income

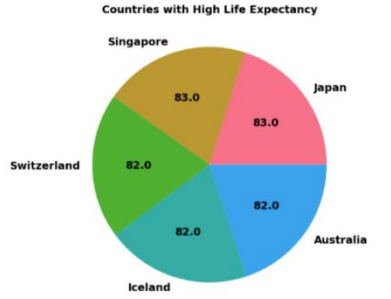


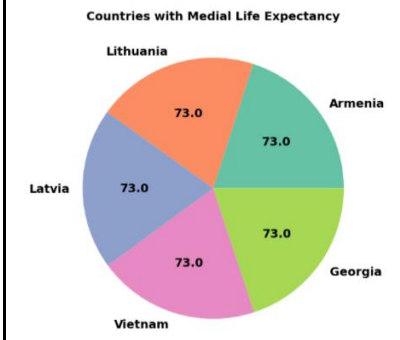


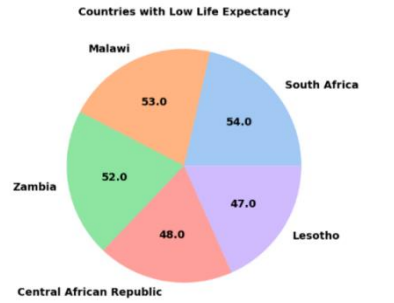


- *Qatar has the highest per capita income. The countries of Luxembourg and Singapore are again seen in the top five countries. In general, South Asian countries are unique in this matter.*
- *African countries are again in the last five lists and their annual income is less.*

d) Life Expectancy

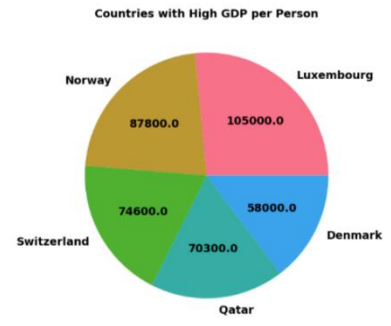


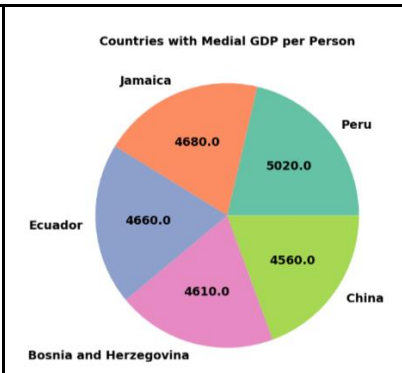


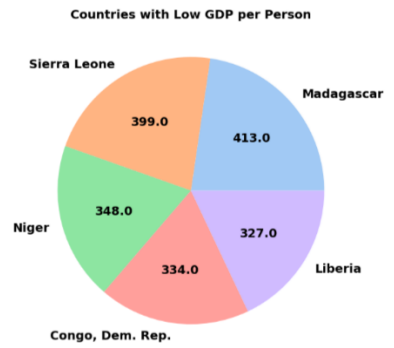


- *African countries have the lowest life expectancy and are again in the last five countries.*

e) Gdp

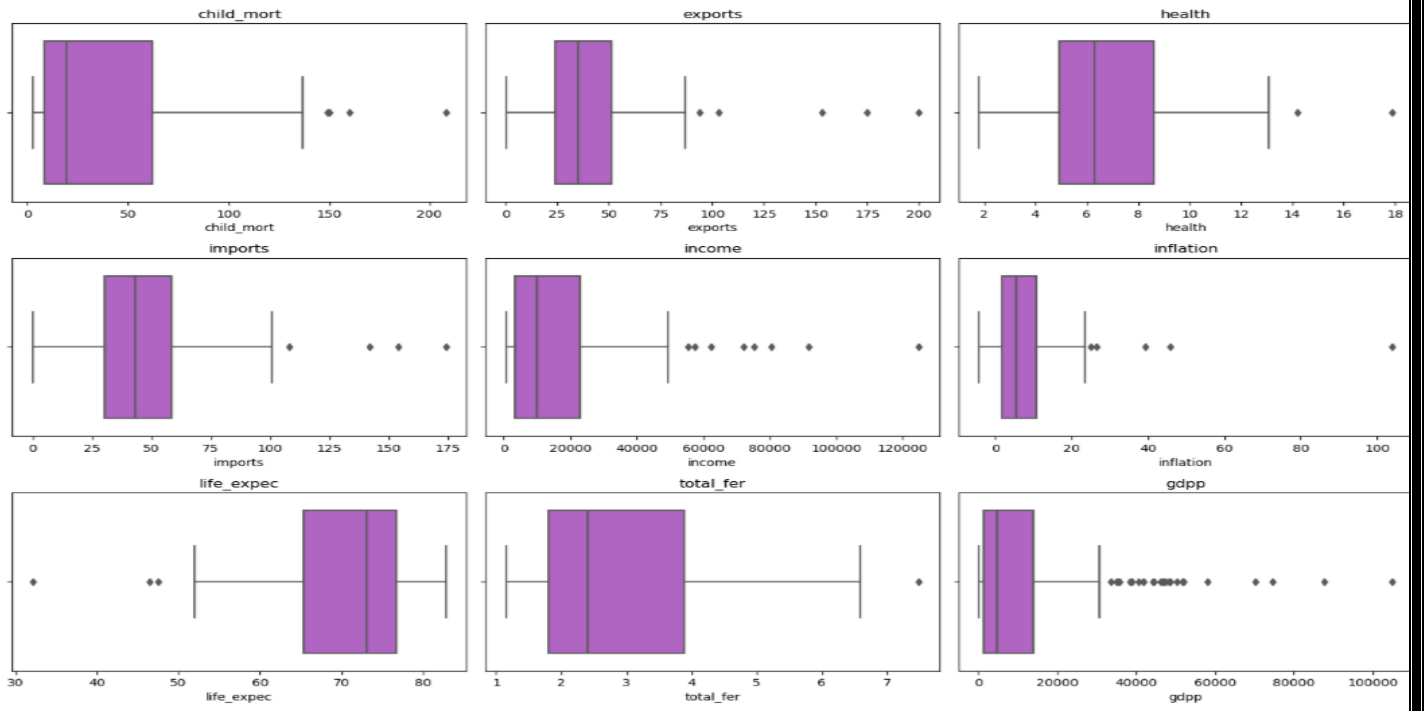






- *European countries have the highest GDP. Luxembourg is again among the top five countries and has the highest GDP. On the other hand, again African countries are on the list of the last five countries.*

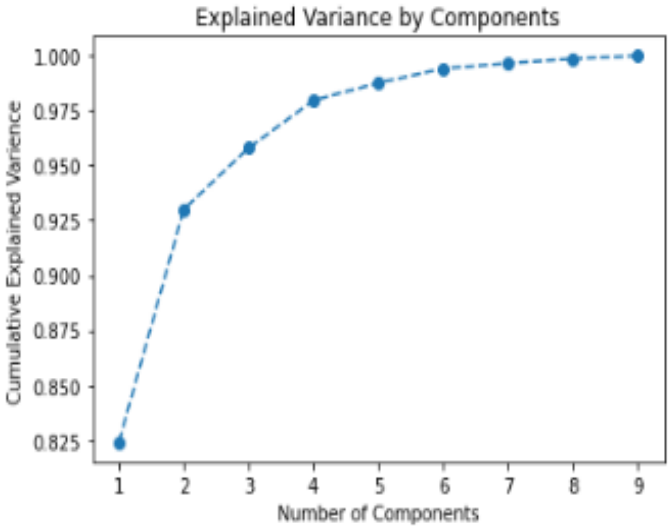
Visualize Noise and Outliers (BoxPlots)

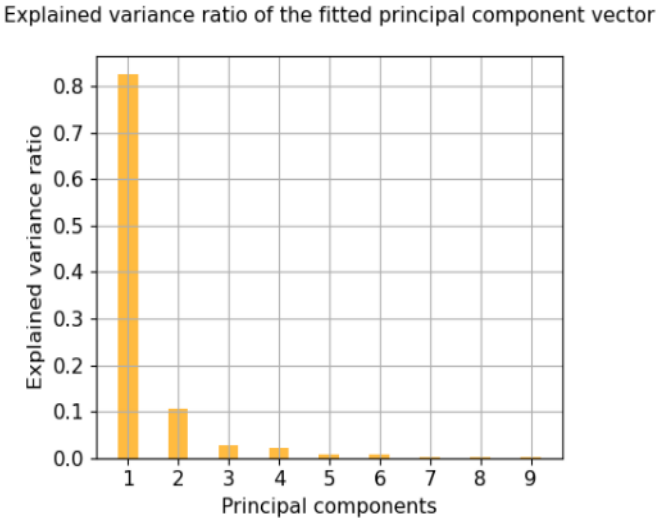


Data don't have outliers or noise. All numbers are possible in the real world!

Dimensionality Reduction

PCA





Optimal n_components = 2

Comparative Study

K Means (num of clusters = 3)					
Standard Scaling		Min Max Scaling		Standard + Min Max Scaling	
No PCA	PCA	No PCA	PCA	No PCA	PCA
0.283	0.442	0.342	0.5	0.359	0.412

The Silhouette score corresponding to K Means implemented on data scaled using standard scaler and reduced using PCA has been ignored as the frequency of number of countries in one cluster was very less and the countries were majorly clustered in only 2 clusters, indicating that the cluster formation was wrong.

DBSCAN															
	Standard Scaling					Min Max Scaling					Standard + Min Max Scaling				
	k (num of clus)	esp (radius)	min samples	metric	Silhouette score	k (num of clus)	esp (radius)	min samples	metric	Silhouette score	k (num of clus)	esp (radius)	Min samples	metric	Silhouette score
No PCA	2	6.45	41	euclidean	0.63	2	1.8	60	manhattan	0.554	2	2.75	60	manhattan	0.644
	3	3.65	16	manhatten	0.235	3	0.65	19	manhattan	0.307	3	0.8	9	manhattan	0.275
PCA	2	6.35	50	manhatten	0.608	2	0.5	7	euclidean	0.554	2	2.3	60	manhattan	0.644
	3	1.25	18	manhatten	0.202	3	-	-	-	-	3	0.5	24	euclidean	0.326

In case of DBSCAN implemented using 2 clusters, we observed that the frequency of number of countries in one cluster was very high while that in another cluster was too low. So, we implemented using 3 clusters, and found out the clustering to be appropriate. For this reason, Silhouette scores corresponding to number of clusters = 2 have been ignored.

Hierarchical Clustering

Hierarchical Clustering(num of clusters = 3)																							
Standard Scaling								Min Max Scaling								Standard + Min Max Scaling							
'ward'		'single'		'average'		'complete'		'ward'		'single'		'average'		'complete'		'ward'		'single'		'average'		'complete'	
No PCA	PCA	No PCA	PCA	No PCA	PCA	No PCA	PCA	No PCA	PCA	No PCA	PCA	No PCA	PCA	No PCA	PCA	No PCA	PCA	No PCA	PCA	No PCA	PCA	No PCA	PCA
0.245	0.43	0.568	0.561	0.562	0.378	0.29	0.433	0.316	0.423	0.425	0.552	0.408	0.438	0.382	0.538	0.348	0.436	0.247	0.479	0.454	0.496	0.456	0.84

Hierarchical Clustering(num of clusters = 4)																							
Standard Scaling								Min Max Scaling								Standard + Min Max Scaling							
'ward'		'single'		'average'		'complete'		'ward'		'single'		'average'		'complete'		'ward'		'single'		'average'		'complete'	
No PCA	PCA	No PCA	PCA	No PCA	PCA	No PCA	PCA	No PCA	PCA	No PCA	PCA	No PCA	PCA	No PCA	PCA	No PCA	PCA	No PCA	PCA	No PCA	PCA	No PCA	PCA
0.248	0.352	0,534	0.558	0.496	0.367	0.286	0.363	0.316	0.407	0.403	0.225	0.4	0.483	0.301	0.48	0.23	0.355	0.146	0.442	0.425	0.346	0.279	0.35

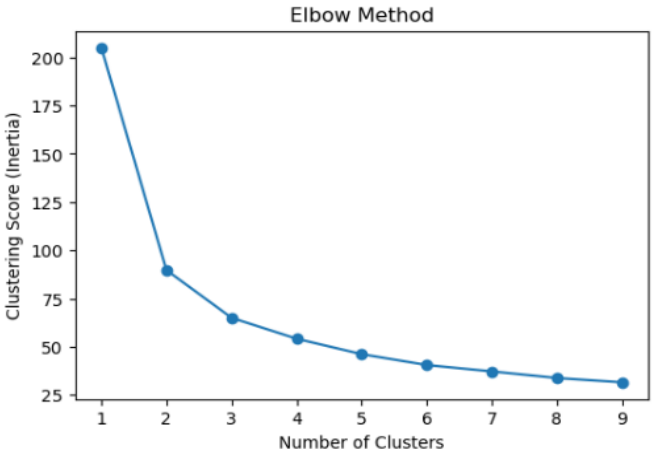
The Silhouette scores corresponding to parameter linkage = single have been ignored as the frequency of number of countries in one cluster was 1 and the countries were majorly clustered in only 2 clusters, indicating that the cluster formation was inappropriate.

Best Models Analysis

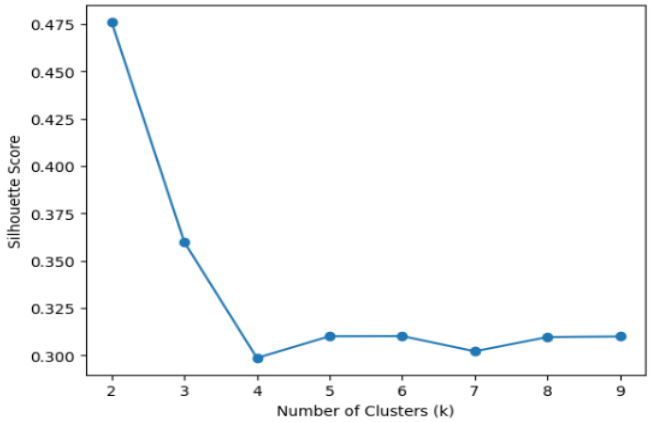
KMeans

For Scaled Data

- **Elbow Method**

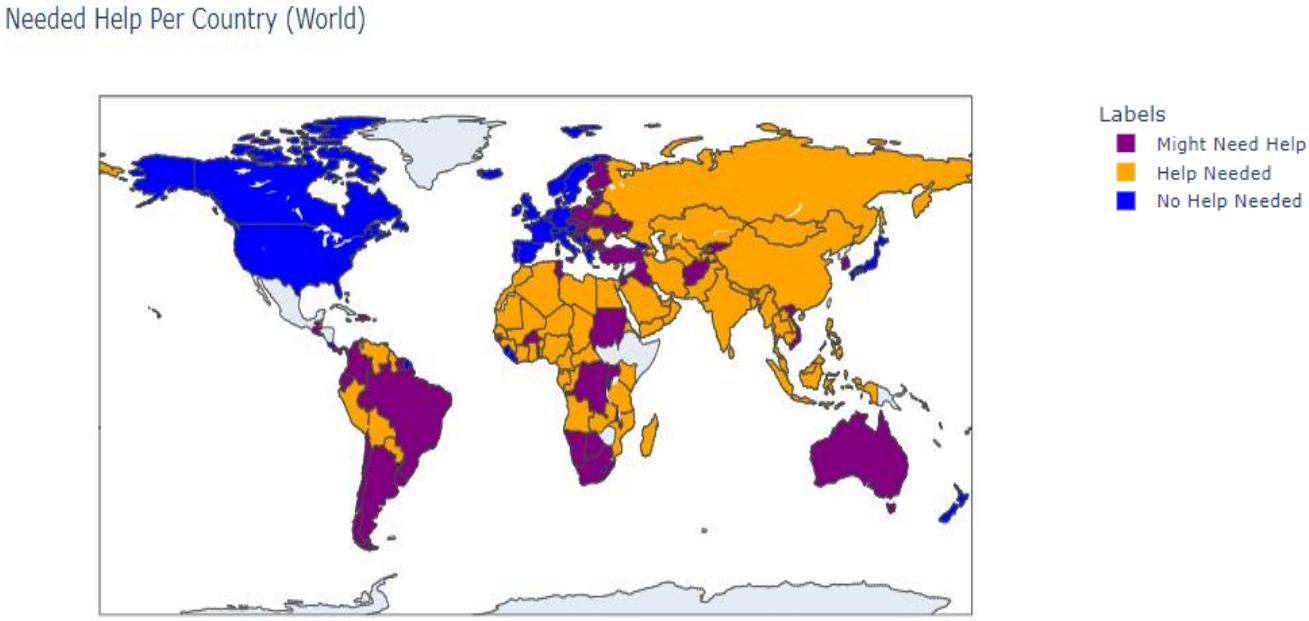


- **Silhouette Score**



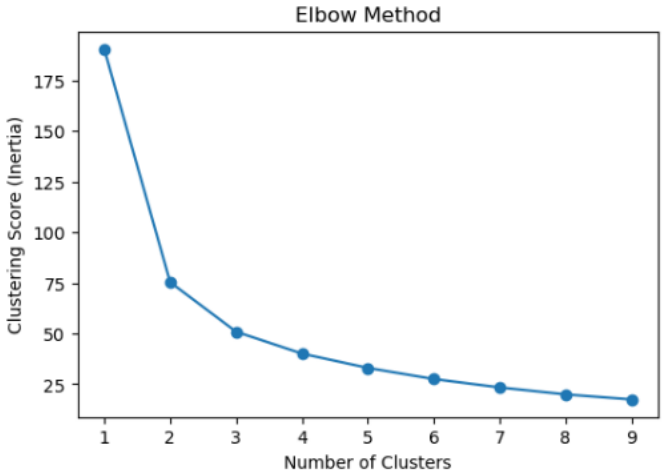

```
k=2, Silhouette score=0.48
k=3, Silhouette score=0.36
k=4, Silhouette score=0.30
k=5, Silhouette score=0.31
k=6, Silhouette score=0.31
k=7, Silhouette score=0.30
k=8, Silhouette score=0.31
k=9, Silhouette score=0.31
```

Best KMeans on Scaled data

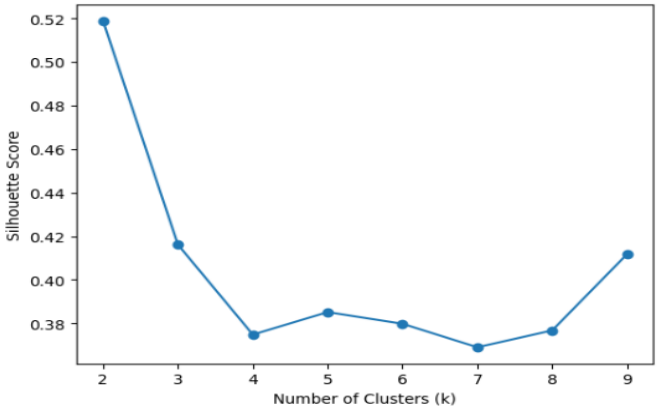


For PCA Data

- **Elbow Method**



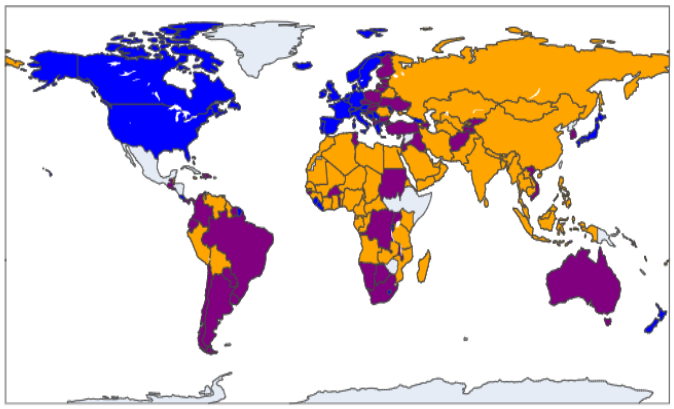
- **Silhouette Score**




```
k=2, Silhouette score=0.52
k=3, Silhouette score=0.42
k=4, Silhouette score=0.37
k=5, Silhouette score=0.39
k=6, Silhouette score=0.38
k=7, Silhouette score=0.37
k=8, Silhouette score=0.38
k=9. Silhouette score=0.41
```

Best KMeans on PCA data

Needed Help Per Country (World)



DBSCAN

Optimal parameters on PCA

N_Cluster	Silhouette_Coefficient
2	0.644741
3	0.326849

For k = 2

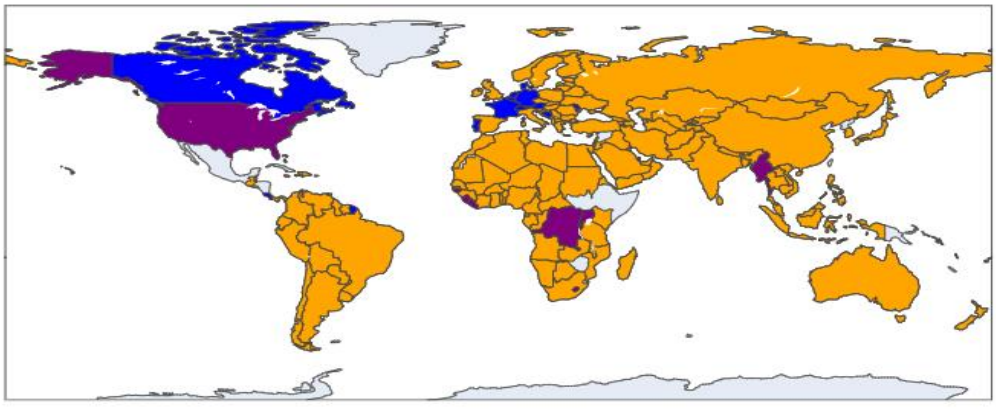
	Distance	Radius	Min_Points	N_Cluster	Silhouette_Coefficient
12711	manhattan	2.3	60	2	0.644741

For k = 3

	Distance	Radius	Min_Points	N_Cluster	Silhouette_Coefficient
19	euclidean	0.5	24	3	0.326849

Best Dbscan model

Needed Help Per Country (World)

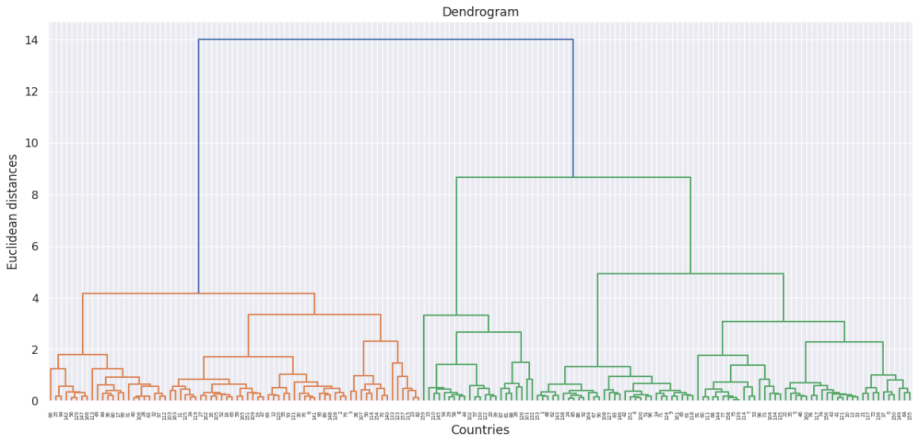


Hierarchical Clustering

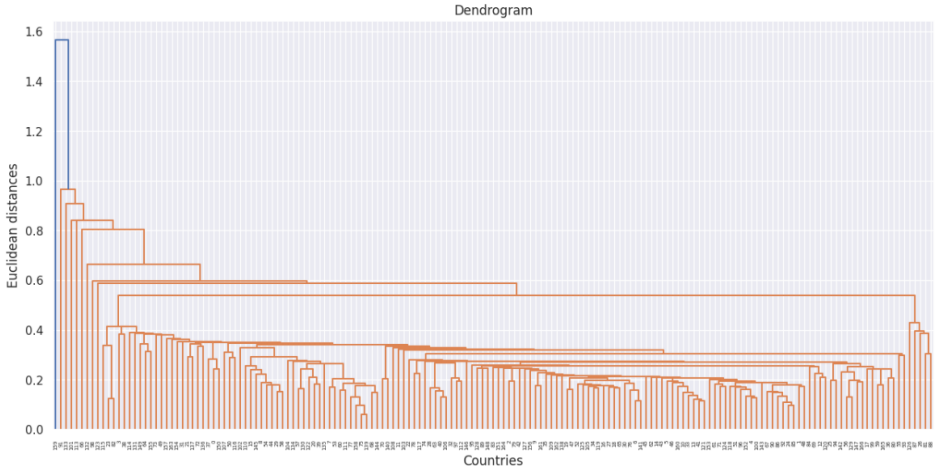
Dendrograms

For scaled

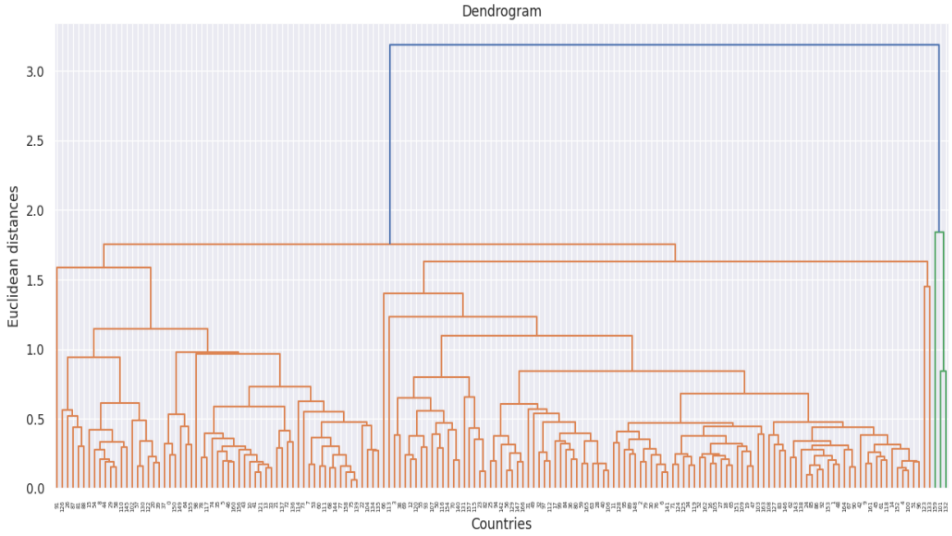
- **Linkage = ‘ward’**



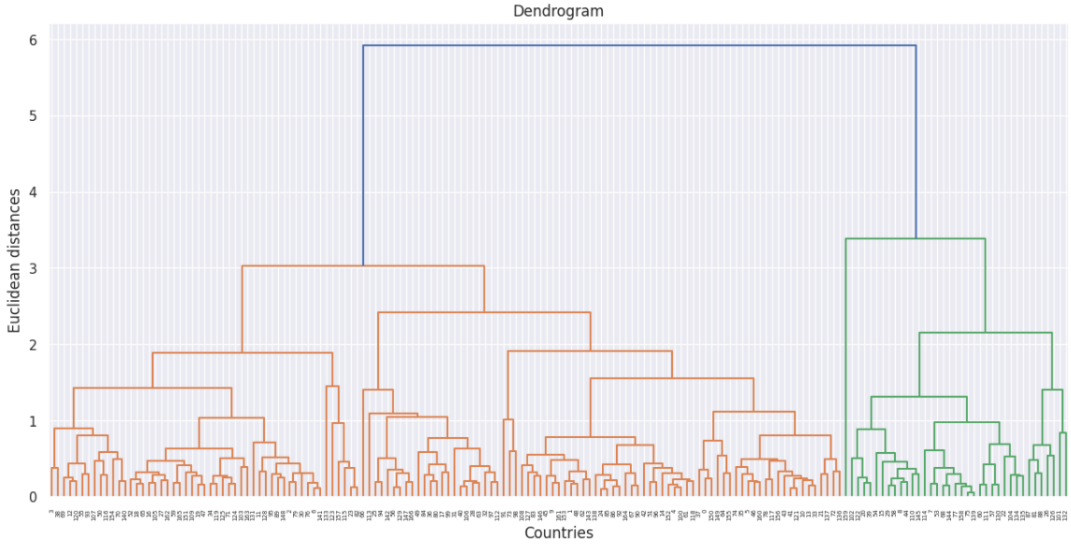
- **Linkage = ‘single’**



- **Linkage = ‘average’**

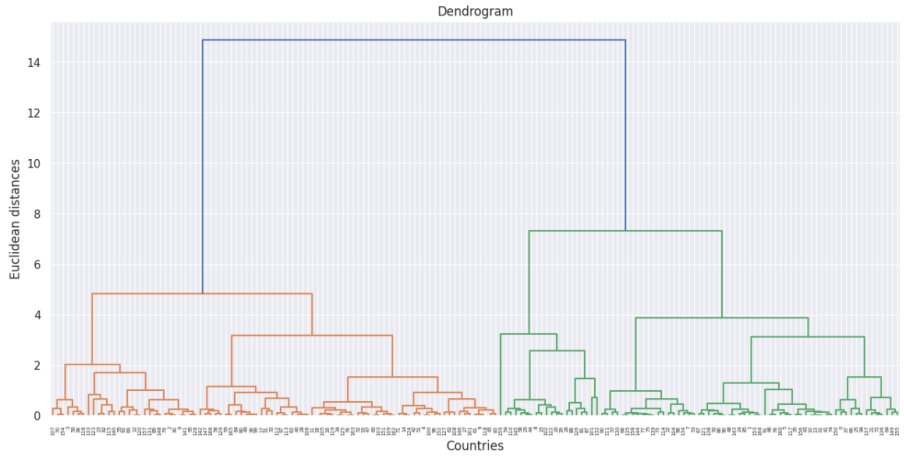


- **Linkage = ‘complete’**

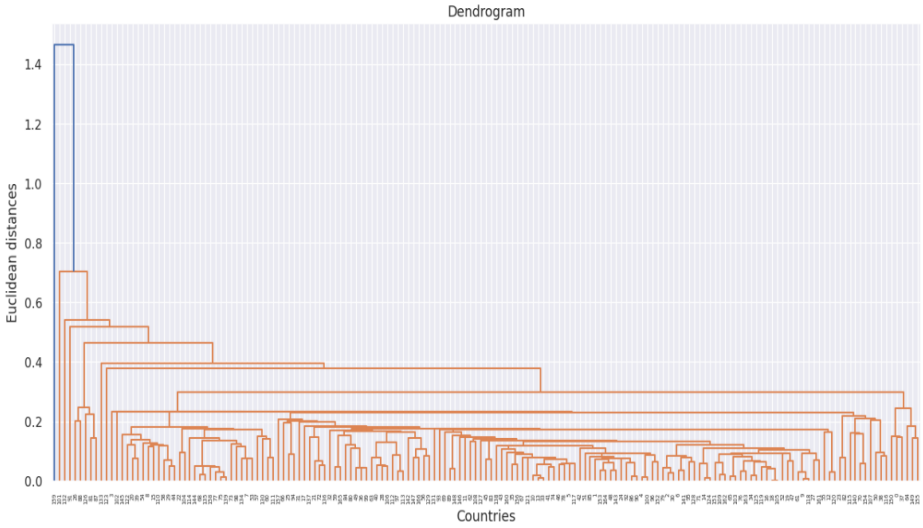


For PCA data

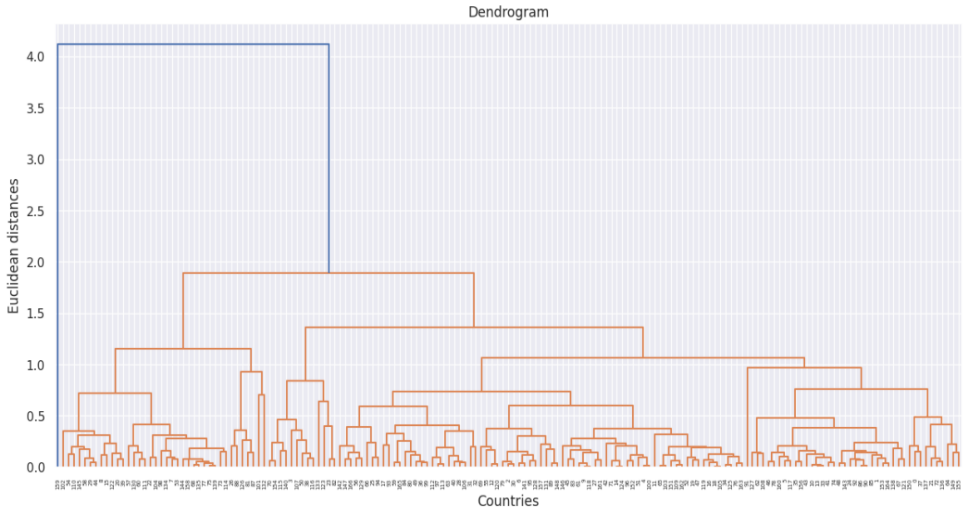
- **Linkage = ‘ward’**



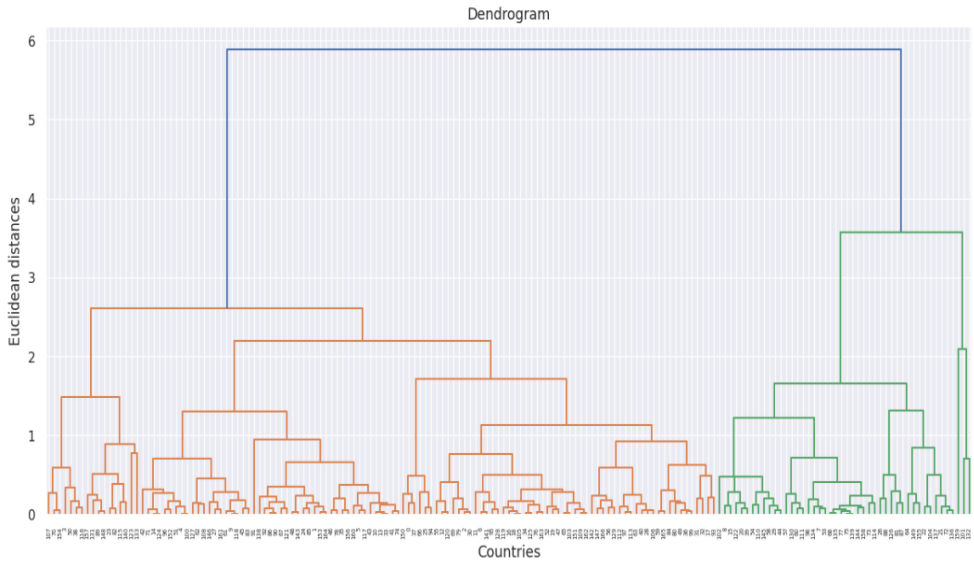
- **Linkage = ‘single’**



- **Linkage = ‘average’**

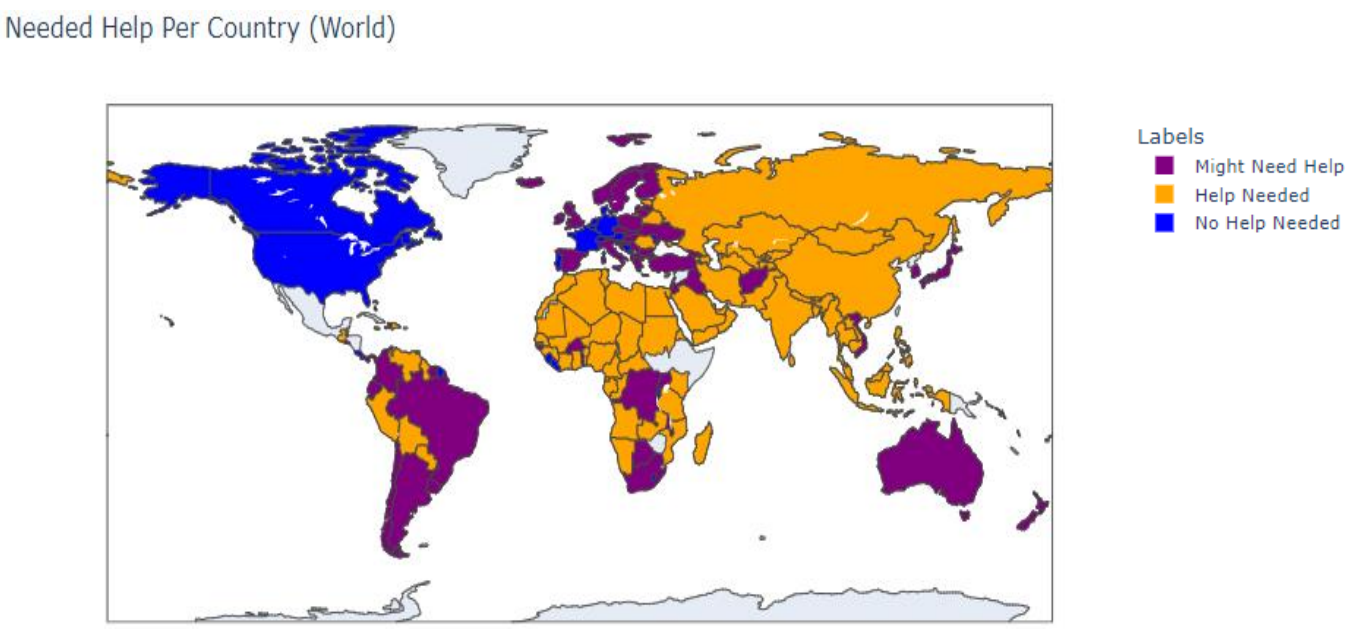


- **Linkage = ‘complete’**



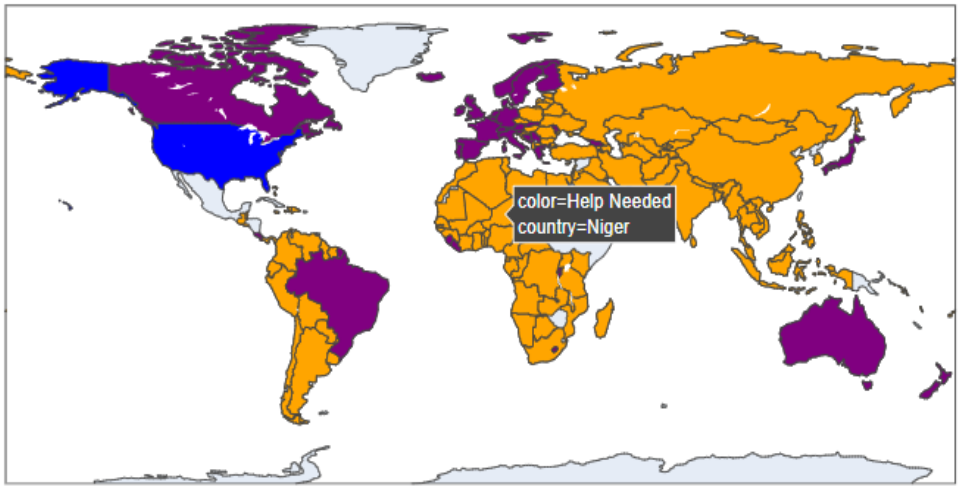
Best Hierarchical Clustering on PCA data

- **Linkage = ‘ward’**



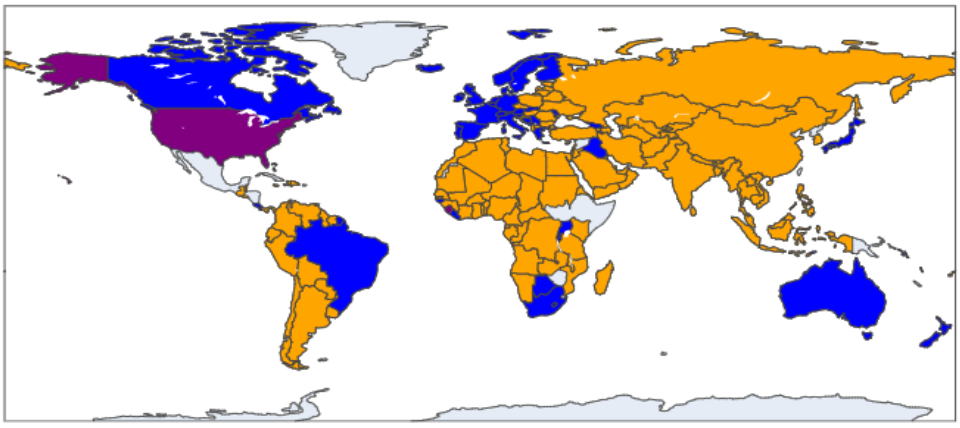
- **Linkage = ‘average’**

Needed Help Per Country (World)



- **Linkage = ‘complete’**

Needed Help Per Country (World)



Conclusion

DBSCAN

We notice that clusters obtained from DBSCAN are not as efficient as K Means & Hierarchical.

For k=2, we see that highly imbalanced data is formed. For k=3, a large number countries are clustered for HELP.

Hierarchical Clustering

From the dendrograms, we find that models using linkage='single' have highly imbalanced clusters with only 1 country in 2 of its clusters. Also, for the PCA data, we find models using linkage='average' are as well inefficient.

Comparing the results for k=3

We choose top 5 models based on silhouette score and do the further analysis. Using PCA Data, we get linkage='complete' as the the most efficient model comparing with linkage='ward', linkage='single' & linkage='average'

K Means

After comparative study, we obtain 2 K Means models, where we use the combinations of standard and minmax scaling.

Best Model

The cluster of Countries that needs money from HELP International NGO because the model is predicated to have less income and health report, and child_mort value seems to be higher than other 2 clusters.

After a long analysis, we obtain 3 models which give better results, i.e. 2 K Means & 1 Hierarchical(linkage='complete') using combination of feature scaling with k=3

Elbow method:

1)The elbow method is used in conjunction clustering techniques, such as k-means clustering, to help determine the appropriate number of clusters to use in the analysis.

2) In this method, the within-cluster sum of squares (WSS) is calculated for different values of k (the number of clusters). The WSS measures the total sum of the squared distance between each point and its assigned centroid in a cluster

Boxplox:

1) a boxplot provides a visual summary of the distribution of a dataset and helps to identify potential outliers in the data.

2) Outliers can be detected by examining the values that fall beyond the upper and lower whiskers of the boxplot. These values are typically considered to be extreme or unusual values and may warrant further investigation.

Kmeans:

- K-means is an unsupervised machine learning algorithm that aims to group similar data points together by iteratively assigning points to the nearest centroid and updating the centroids based on the assigned points until convergence is achieved, resulting in K distinct clusters.

Hyper-parameters:

1) **Kmeans++:** "K-means++" is an initialization method that aims to choose the initial centroids in a way that improves the convergence rate and the final clustering result. It works by selecting the first centroid randomly from the data points and then selecting the remaining centroids from the data points with a probability proportional to their distance from the already chosen centroids

2) **Random:** “Random" initialization method randomly selects K data points from the dataset as the initial centroids. This approach is simpler and faster than K-means++ but may not necessarily produce good initial centroids for all datasets, leading to slower convergence and poorer clustering results.

3) **Hierarchical clustering**: Hierarchical clustering is an unsupervised machine learning algorithm used for grouping similar data points into clusters.

2) The algorithm starts with each data point as a separate cluster and iteratively merges the two closest clusters based on a distance metric until all data points belong to one cluster

3) The output of hierarchical clustering is a dendrogram, which is a tree-like diagram that shows the hierarchy of the clusters.

Linkage: Linkage refers to the method used to measure the distance between clusters in hierarchical clustering.

1) **Single linkage:** This method calculates the distance between the closest two points in each . Ward linkage: This method minimizes the variance of the clusters being merged. It tends to form compact, spherical clusters.

2) **Ward linkage:** This method minimizes the variance of the clusters being merged. It tends to form compact, spherical clusters.

3) **Average linkage:** This method calculates the average distance between all pairs of data points in each cluster and merges the clusters with the smallest average distance. It can produce clusters of varying shapes and sizes.

The choice of linkage method can impact the final clustering result, as each method has its own strengths and weaknesses.

Affinity:

In clustering algorithms, affinity refers to the distance metric used to measure the similarity or dissimilarity between data points

1) **Euclidean distance:** This metric is the straight-line distance between two data points in a multi-dimensional space**.**

2) **Manhattan distance:** This metric is the sum of the absolute differences between the coordinates of two data points in a multi-dimensional space.

DBSCAN:

1) Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is an unsupervised machine learning algorithm used to cluster together points in a dataset based on their density.

2) The algorithm works by grouping together points that are within a certain distance of each other and have a minimum number of neighbours. Points that do not meet these criteria are considered outliers

3) The resulting clusters can have any shape and can vary in size, making DBSCAN a useful tool for exploratory data analysis and anomaly detection.

COMPARISONS BETWEEN KMeans, Hierarchichal, and DBSCAN clustering

Algorithm	Type	Advantages	Disadvantages
K-Means	Centroid-based	1. Simple and computationally efficient.	1. Sensitive to the initial placement of centroids.
		2. Good for high-dimensional data.	2. Requires a predetermined number of clusters.
		3. Fast convergence with large datasets.	3. May converge to suboptimal solutions.

DBSCAN	Density-based	1. Can find clusters of arbitrary shape.	1. Sensitivity to hyperparameters.
		2. Can identify noise points.	2. Can be computationally expensive.
		1. No need to specify the number of clusters.	
Hierarchical Clustering	Connectivity-based or Agglomerative	1. No need to specify the number of clusters.	1. Can be computationally expensive for large datasets.
		2. Density-based	2. Sensitive to the choice of linkage method.

DIFFERENCE BETWEEN PCA, LDA, AND ICA

Algorithm	Purpose	Method	Output
PCA	Dimensionality reduction and feature extraction	Linear transformation	Reduced dimensionality dataset
LDA	Feature extraction and classification	Linear transformation	Discriminant features
ICA	Blind source separation and feature extraction	Nonlinear transformation	Independent components

Reference Links-

- kmeans-sklearn
- Hierarchical-sklearn
- DBSCAN-sklearn
- Nearest Neighbors-sklearn
- Silhouette score-sklearn
- Calinski-sklearn
- PCA-sklearn
- ICA-sklearn
- LDA-sklearn