Clustering and PCA- Assignment

SUBMISSION

Ramya sri sai Ch

Abstract

About:

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.

Objective :

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.

Analysis

Procedure :

Steps involved:

1. Perform PCA on the dataset and obtain the new dataset with the Principal Components. Choose the appropriate number of components k.
2. Apply K-means clustering to form clusters on the data set.
3. Apply Hierarchy clustering on the data set obtained after step 1 to create Clusters
4. Analyze the clusters and identify the ones which are in dire need of aid.

Data Preparation

1.Checking if there are null values in the data.

2.Convert the columns which are described in the form of percentage to absolute values.

3.Check if duplicates are there

EDA :

Perform EDA on the data.

On simply finding out the least income generating countries, high child mortality rate, life_exp does not give results directly so we are performing PCA as there is high correlation between the variables.

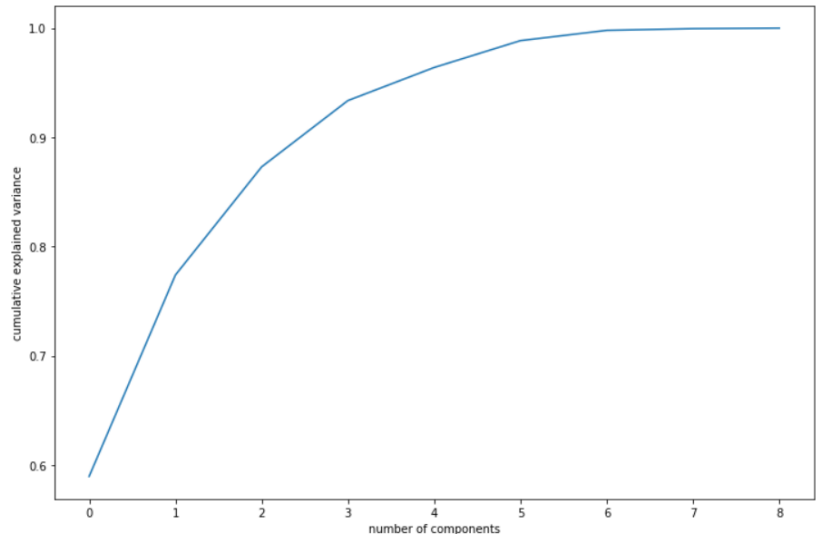
Principal Component Analysis

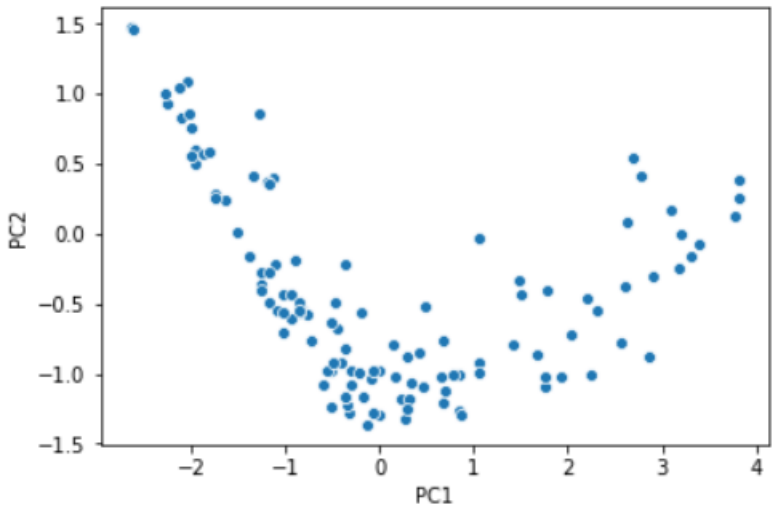
Scaling of variable by standard scalar method

Applying PCA on the data

find out the pricipal component

screen to plot to find the number of pcs, from the graph the best values for components are 3 where 90% of variance is being explained.





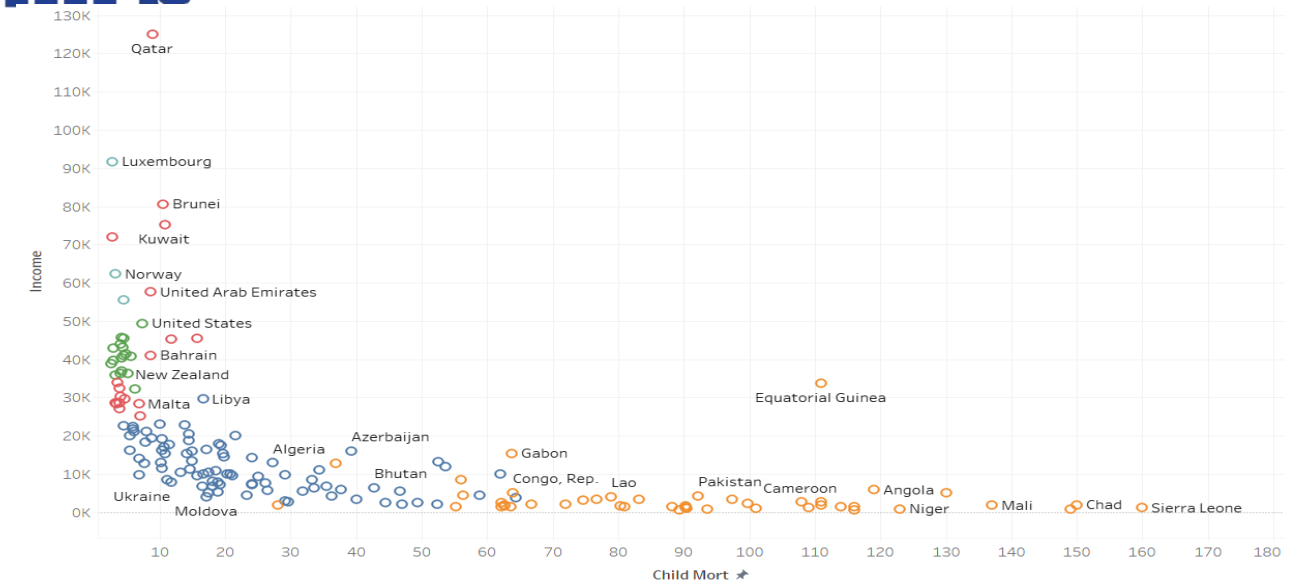
Remove outliers in the data if there in the data

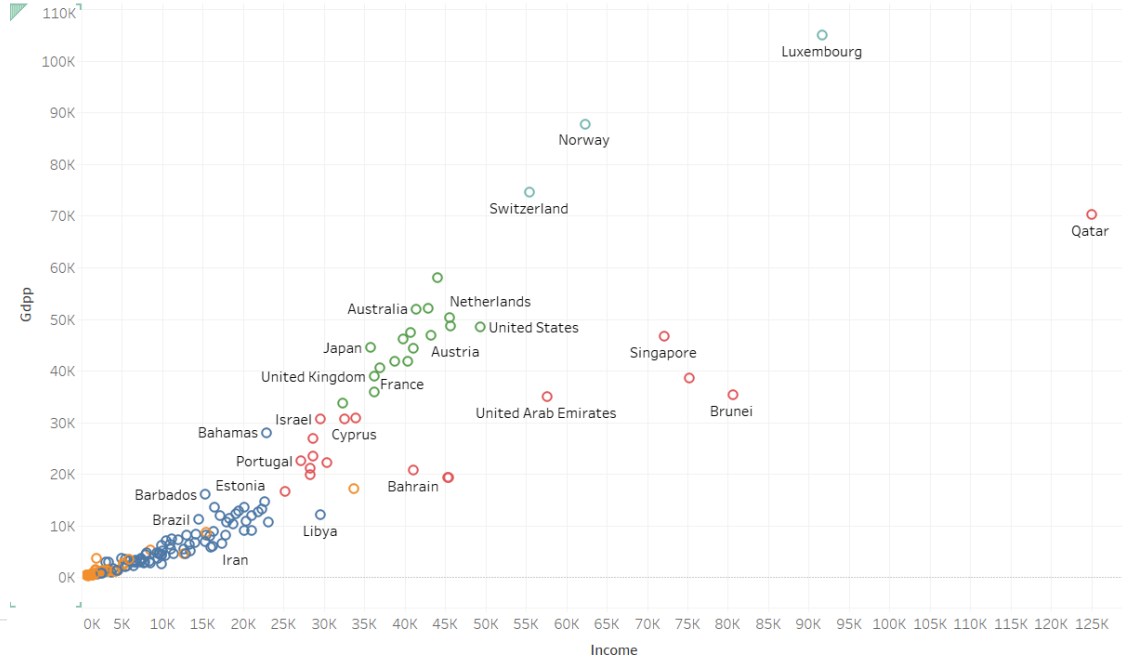
The scatter plot of pc1 and pc2 show how the data is spread

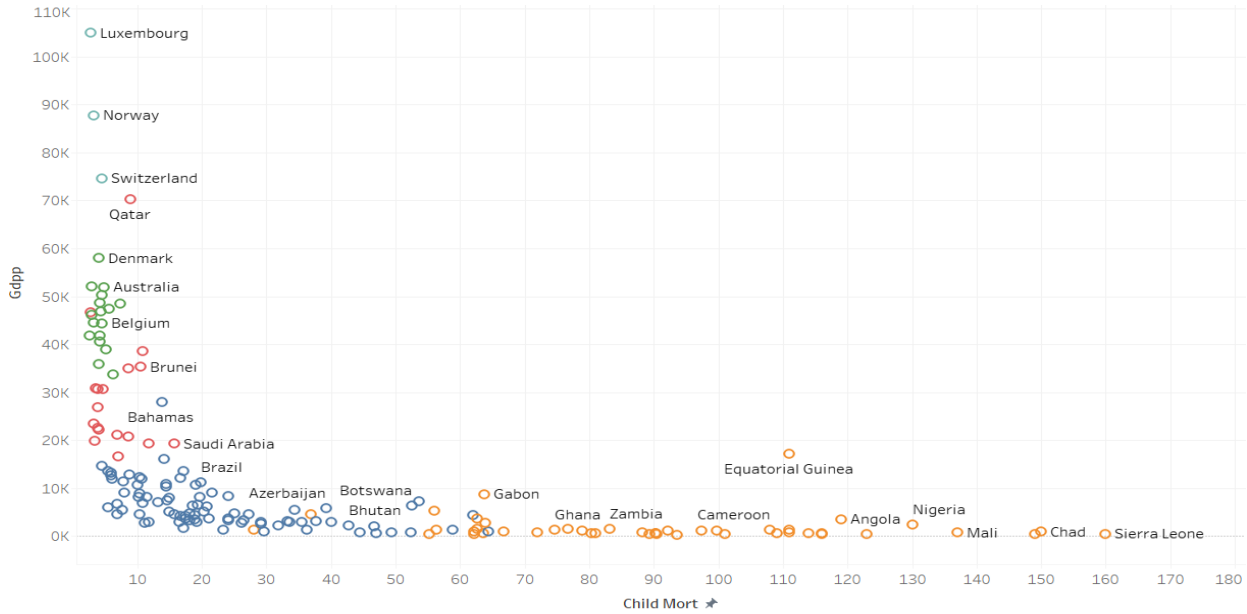
Calculate the Hopkins statistic to ensure that the data is good for clustering.

K-Means

1. To identify the number of clusters that are being formed is need to be checked by the silhouette score.
2. Also use elbow curve to find the k values to identify the number of cluster that can be formed



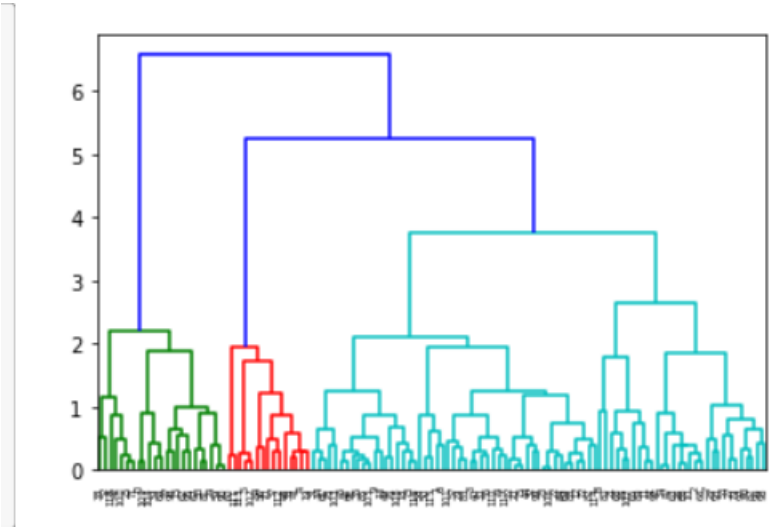




The Clusters are being formed between variable .

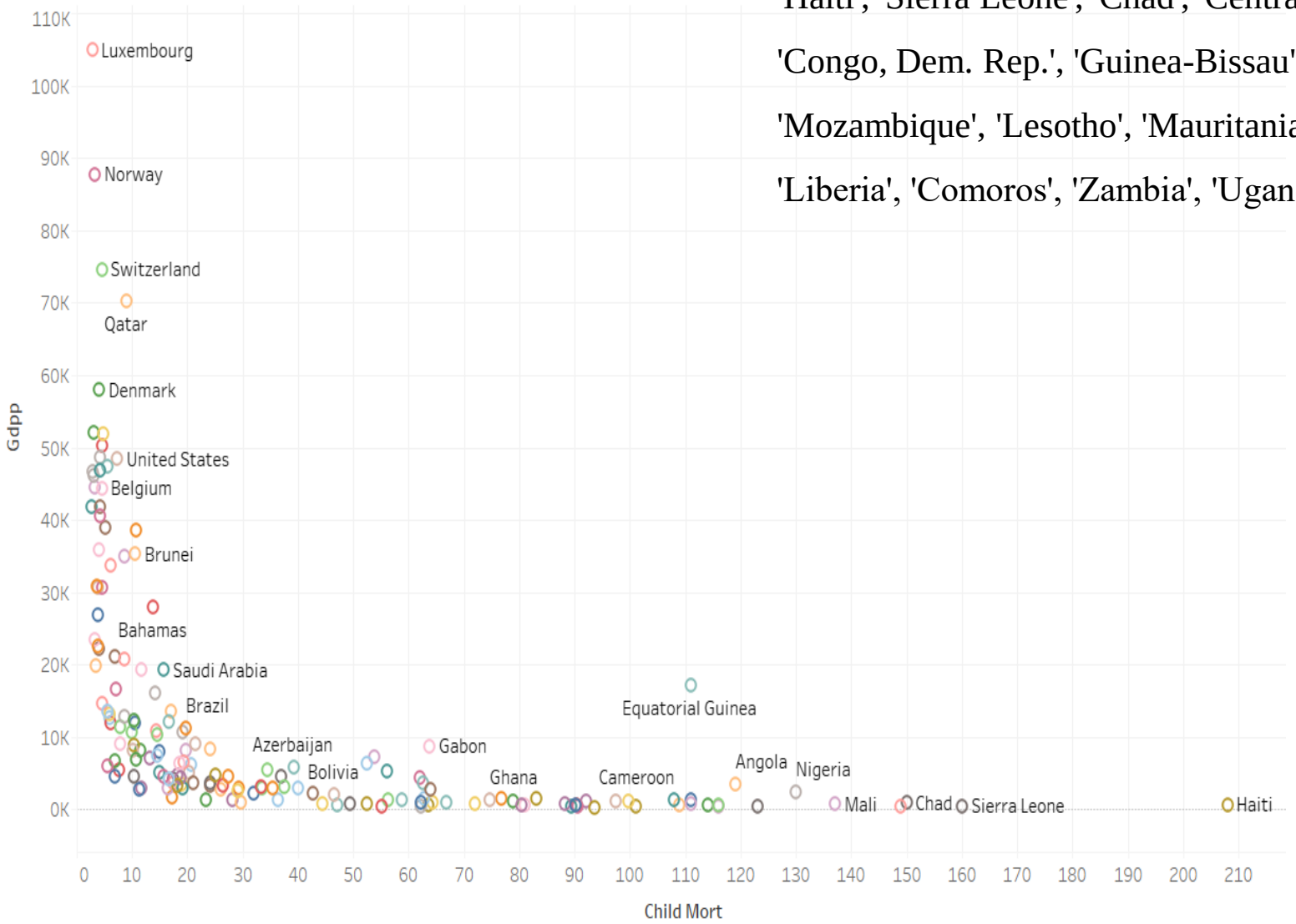
If we see clearly we can found the countries which are in dried need belong to same cluster

Hierarchical Clustering



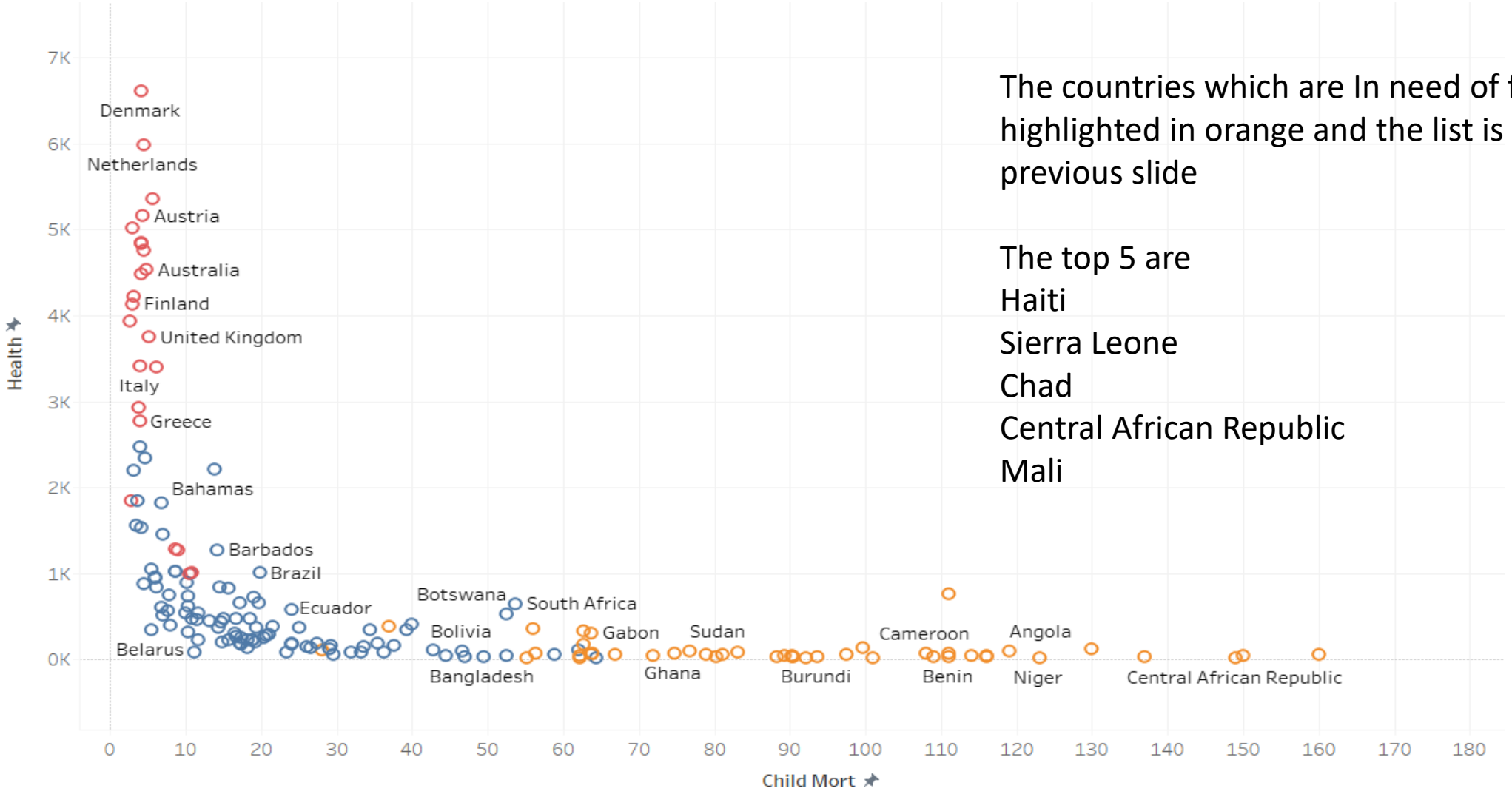
Hierarchical Clustering is also one of the method to form clusters. Here k need not to be pre defined

Results



The countries after clustering are found out to be funded are:

'Haiti', 'Sierra Leone', 'Chad', 'Central African Republic', 'Mali', 'Niger', 'Burkina Faso', 'Congo, Dem. Rep.', 'Guinea-Bissau', "Cote d'Ivoire", 'Benin', 'Guinea', 'Cameroon', 'Mozambique', 'Lesotho', 'Mauritania', 'Burundi', 'Pakistan', 'Malawi', 'Togo', 'Afghanistan' 'Liberia', 'Comoros', 'Zambia', 'Uganda', 'Gambia', 'Lao', 'Sudan', 'Ghana', 'Tanzania‘



The countries which are In need of funding are highlighted in orange and the list is specified in the previous slide

The top 5 are
Haiti
Sierra Leone
Chad
Central African Republic
Mali