Question 1: Assignment Summary

Approach

- **Data Collection and Data Cleaning:**

Importing the data then cleaning, checking if there are any null values.

- **Visualising data:**

We detected outliers by visualising the data, outliers were treated according to our problem statement. Also found out that some Variables are highly corelated to each other.

- **Outliers Detection and treatment**

There were outliers in almost every column. For this analysis we capped the gdppcolumn at .95 quantile at upper end and .05 quantile at lower end.

- **Scaling data**

Standardizing all the continuous variables.

- **Hopkins test:**

To check if data has tendency to form clusters.

- **Kmeans Clustering:**

Identifying the “k” through silhouette analysis and elbow curve. Then forming the cluster on scaled data then adding the cluster id on original data for better interpretation of data. And then visualizing the clusters.

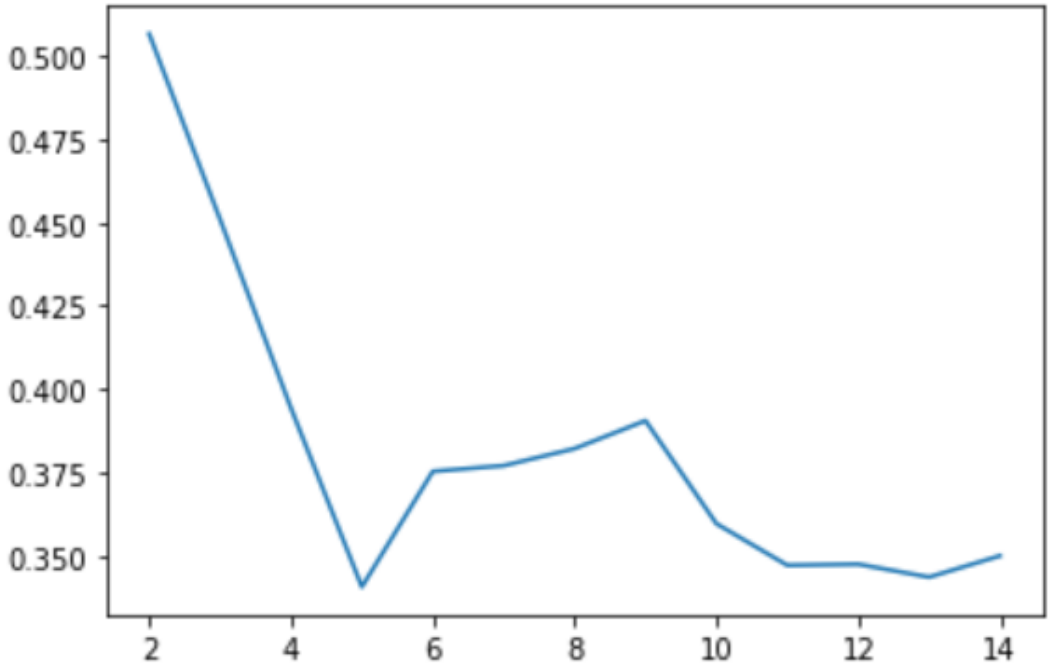
- **Hierarchical Clustering**

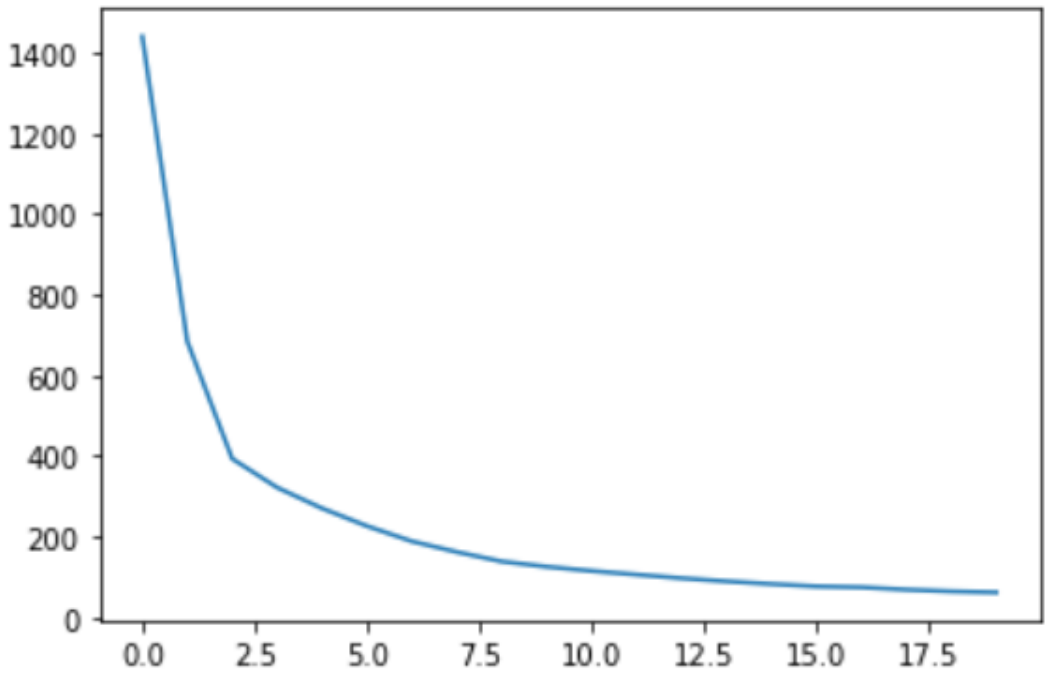
Identifying optimal number for k by analysing dendrogram. Then forming the cluster on scaled data and adding the cluster label to original data for better interpretation. Visualisation of clusters was also done.

- **Decision Making:**

Successfully identified the countries by analysing both model which are in dire need of Aid.

Silhouette Analysis





The elbow curves at around 4 on X-axis, hence we will choose 4 clusters for K means

Summary

We got same top 10 countries from both hierarchical and k-means model that are in dire need of Aid.

Following are the countries name requiring aid.

1. Sierra Leone
2. Central African Republic
3. Haiti
4. Chad
5. Mali
6. Nigeria
7. Niger
8. Angola
9. Congo , Dem. Rep.
10. Burkina Faso