Clustering Assignment

Presented By:

K Lakshmikanth

Task Performance on Dataset

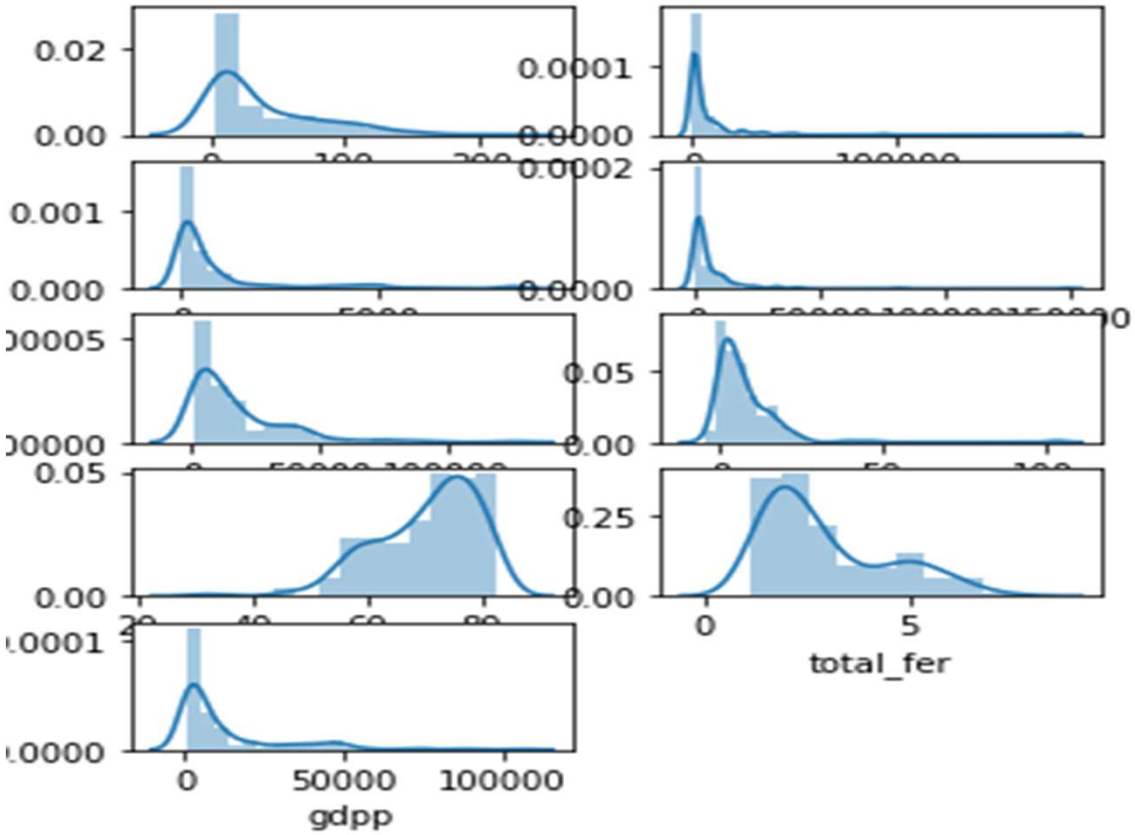
Task-1: Reading and Understanding Data

- Importing all necessary libraries
- Importing .**csv** file:
- Inspect the structure of the data
  - df.shape
  - df.describe( )
  - df.info( )
- Inspect Null values (both in columns and rows of Dataframe)

Task-2: Changing the Units of Columns

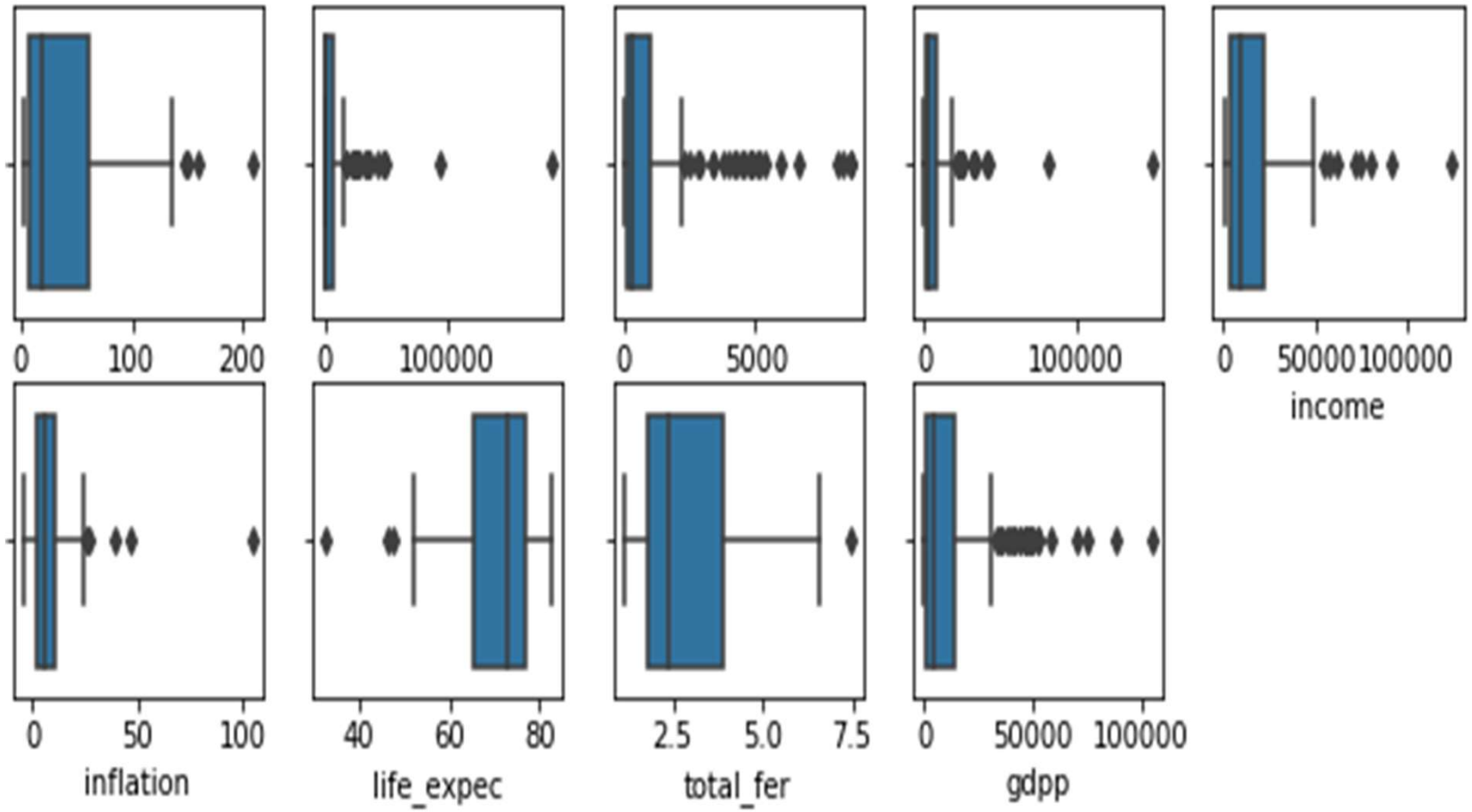
- Checking all columns
- Converting "exports", "imports" and "health" of spending percentages.
- Inspecting the dataframe after converted columns

- Hence this is unsupervised learning(Clustering) we don't the label i.e, "country". So, we are dropping "country"
- Confirming the new data frame which doesn't have the column "country“
- Inspecting the data distribution of various columns
- plotting the data of various columns



Inspecting The Outliers

- Inspecting the outliers by plotting.



Outlier Treatment – Capping

- For lower range outliers, if we cap them we may loose the "country" which are in requirement of the AID. So, we don't cap the lower range outliers.
- incase of "child_mort" we need upper range outliers to analysis the "country" which are in requirement of the AID.
- capping the upper range outliers of "exports“, "health“, "imports“, income“, "inflation“, "total_fer“, "gdpp“
- Checking the statistical information and presence of outliers by describing.

Task-3: Calculating the Hopkins statistic

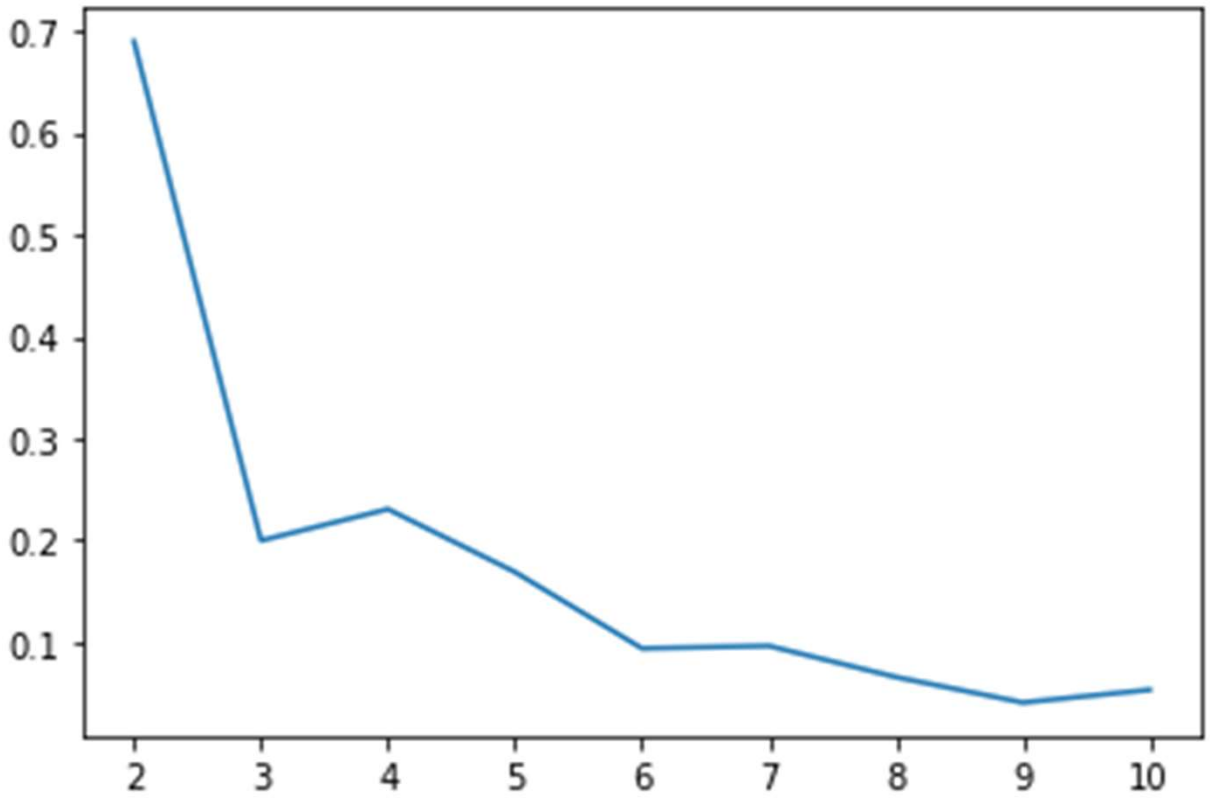
- Calculating the Hopkins statistic
- The mean value of Hopkins statistic is 86% so that the data is good enough for clustering

Task-4: Scaling of data

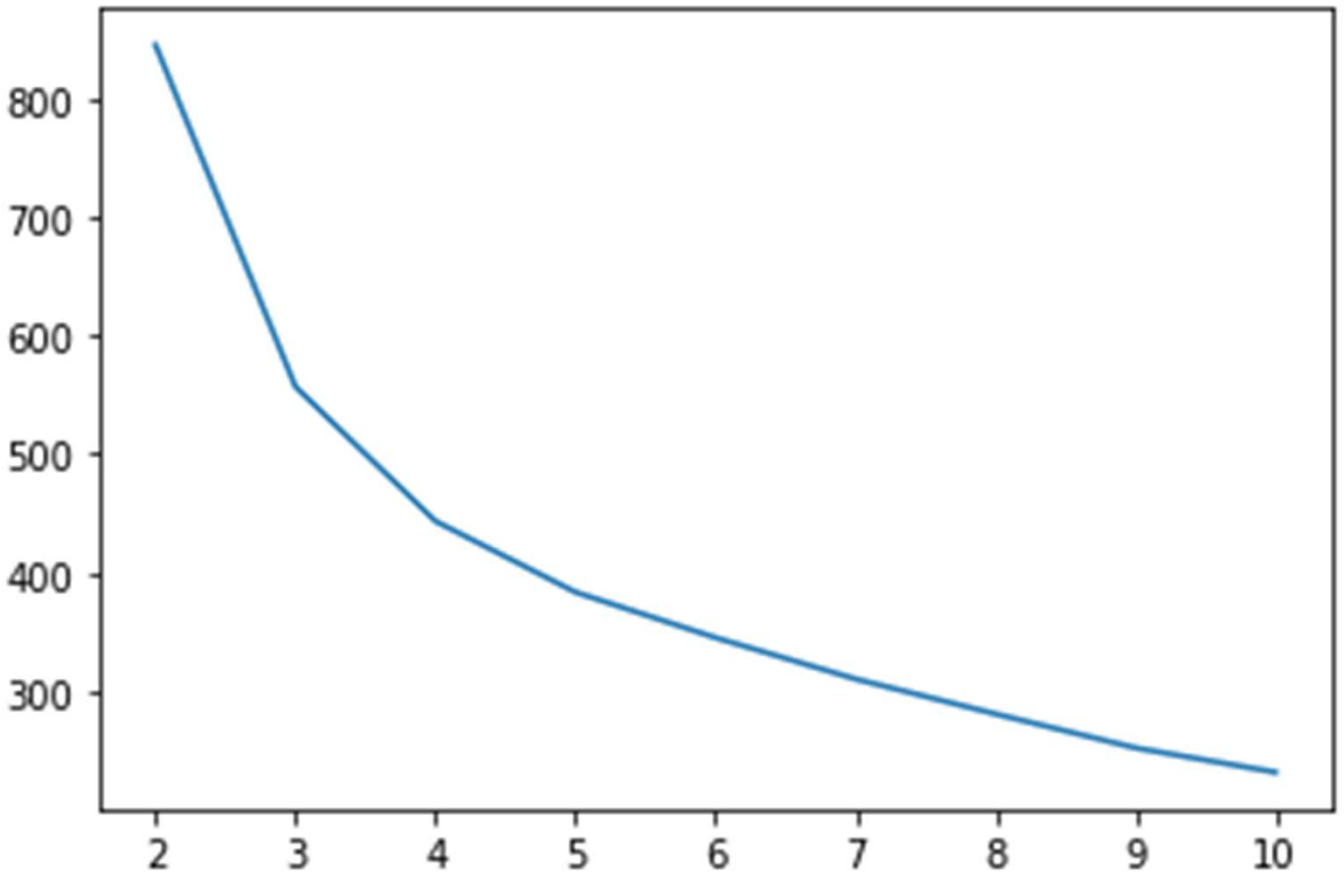
- Values of attributes are in different scales can distort this distance. Hence, we need to bring attributes into the same scale using standardization metric.

Task-5: Clustering

- Kmean Clustering
  - Plotting the ‘Elbow curve’



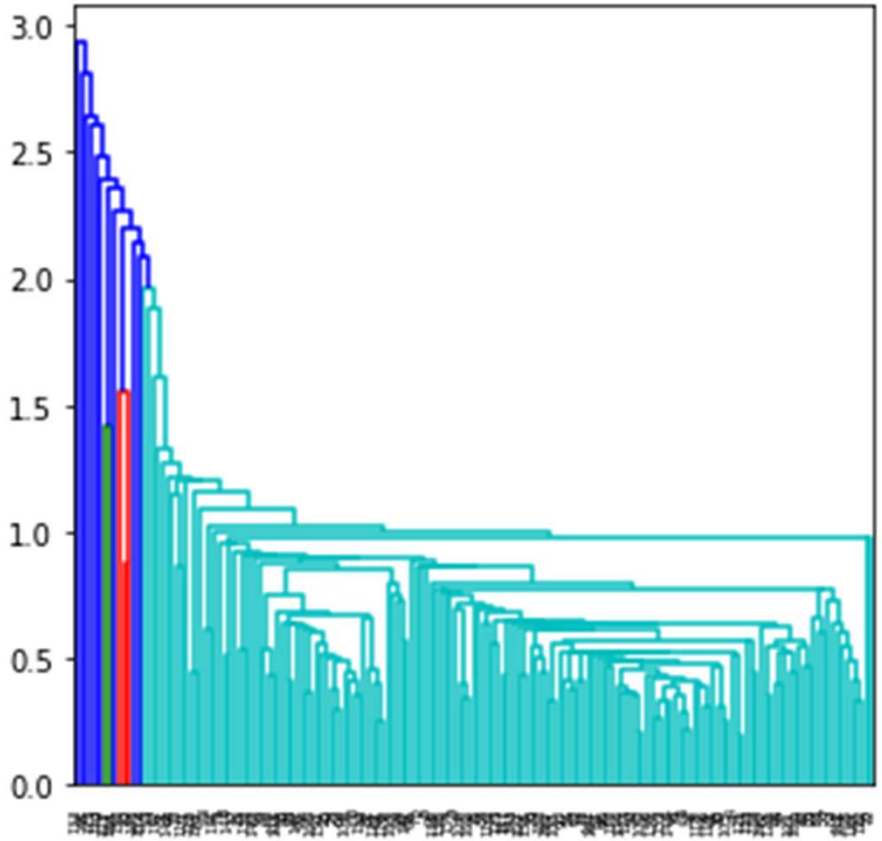
➢Plotting the ‘ssd curve’



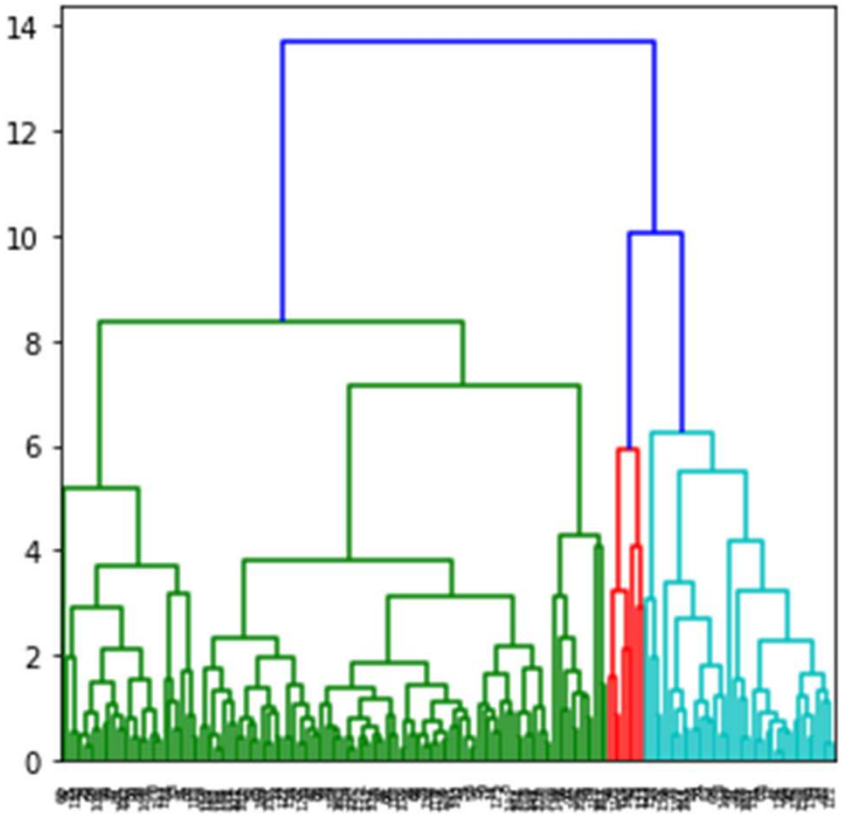
- Hierarchical Clustering
- Single Linkage : The following are observed from the above clustering Profile and Value counts of Single Linkage : Only one cluster is dominating other clusters i.e. cluster label 0 The total count of cluster label 0 is 165 other cluster label is having one count each Hence Single Linkage Cluster is not used in the further analysis
- Complete Linkage:Similar points can be observed as seen in KMeansClusters: Inflation effect on child_mort and Life_expec : except on data point there in no much impact on inflation GDPP effect on Child_mort and Life_expec: higher the GDPP lower the child_mort and higher life_expec Higher spending on health there is a lower child_mort and Higher Life_expec Higher the total_fer, lower in life_expec and higher child_mort

Visualizing the clustering

Single Linkage



Complete Linkage



- **Conclusion :**

As per K_means Clustering (cluster number 1),the following countries require aid by considering the socio – economic factor:

1.Haiti

2.Sierra Leone

3.Chad

4.Central African Republic

5.Mali

As per Hierarchical Clustering (cluster number 0),the following countries require aid by considering the socio – economic factor:

1.Haiti

2.Sierra Leone

3.Chad

4.Central African Republic

5.Mali