CLUSTERING ASSIGNMENT

Identify top - 5 countries that are direst need of AID

Presented By :
V. Prabhaakar

Tasks Performed On Dataset

Task-1:

- **Importing .csv Files:** country-data.csv

SubTask-1.1:

- **Inspect the data frame / Check the structure of the data**
  - converting the %age values "exports,health, imports " into actual values
  - df.shape
  - df.info ( )
  - df.describe()

Task-2

- **Data Quality Check and Missing values/Cleaning the Data**
  - Inspect Null values (both in columns and rows of data frame)

Observation :None of the columns have null values hence not required to drop values.

Task-3: Data Visualisation/Perform EDA to understand various variables

From the plots we can observe following:

- cluster profiling is possible on child_mort,inflation, GDPP,exports,imports, Income , life_expec, Total_fer

Subtask 3.1 : Outlier Treatment

- From plots we can observe the following:
- Upper outliers exist for child_mort, exports,imports,inflation,health, income, total_fer,and GDPP
- Lower Outliers exist in Life_expec As we need to find the direst need of AID, so we should not treat the upper outliers of Child_mort and Inflation.
- For analysis purpose we are treating the upper outliers using Capping
- We can observe that all the columns having upper outliers are capped to 99% except Child_Mort and Inflation

Task-4:

Measuring the cluster tendency (Hopkins Statistics)

Find the cluster tendency

- Hopkins stats run: 100 times The Mean Value of Hopkins is: 0.91
- Consider Data is very good for Cluster

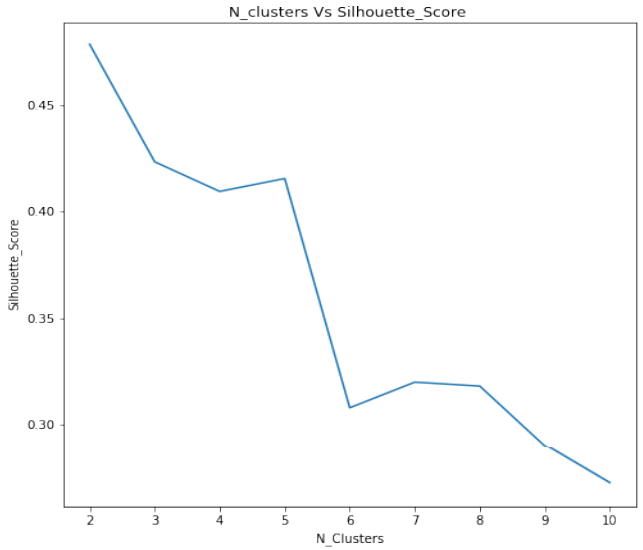
Task-4.1

- **Scaling is performed on the columns**

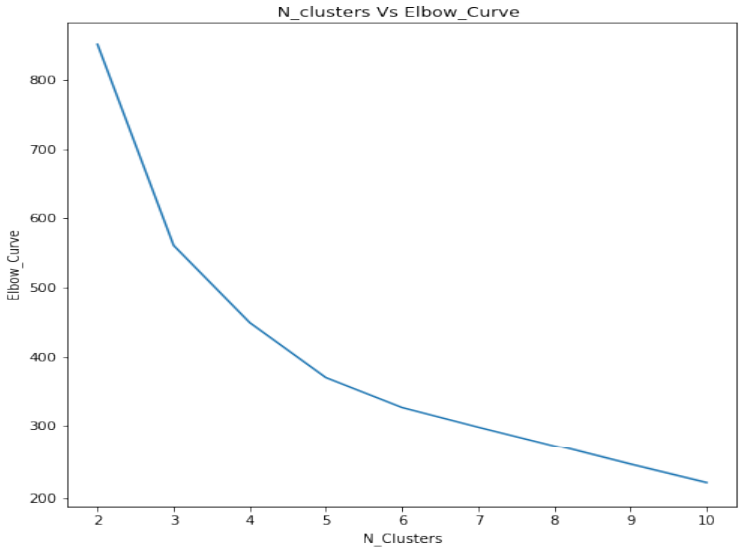
Task-5:

- **Find the K Value used for analysis:**

Silhouette Score



Elbow Curve



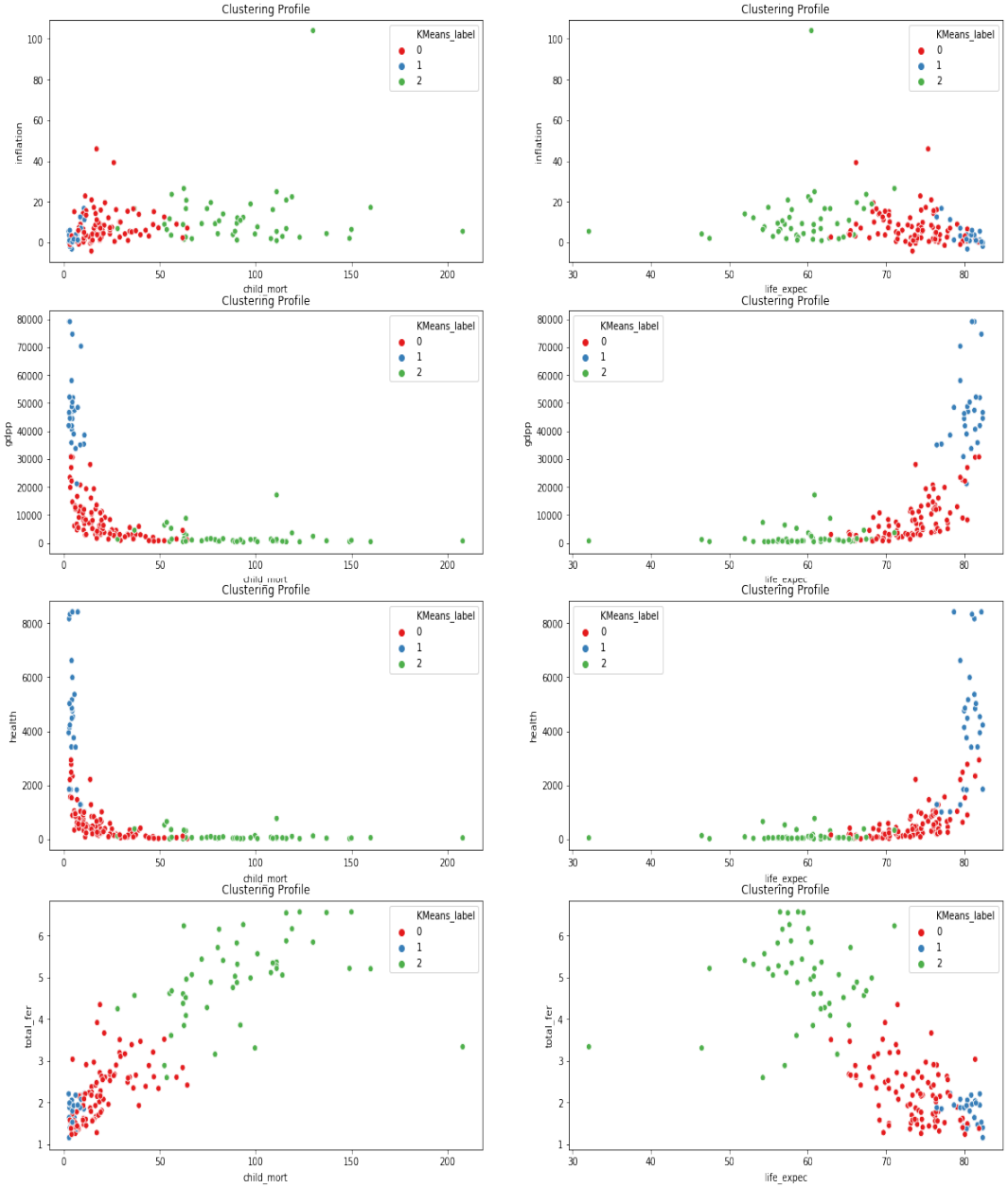
- From the above plots of we can observe the following:
- Silhouette Score for n_clusters 3 (excluding n_clusters = 2) is high as compared to others
- from Elbow curve we can observe the bend at point 3 (n_clusters = 3)
- As 3 is satified in both the plots , we are considering the no of culters as 3 **for analysis**

Task 6: Clustering Profiling using KMeans

SubTask6.1

Cluster Profiling

- Visualization: GDPP - Income, Income - Child_mort, GDPP - Child_mort
- **From the plots we can observe the following:**
- 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

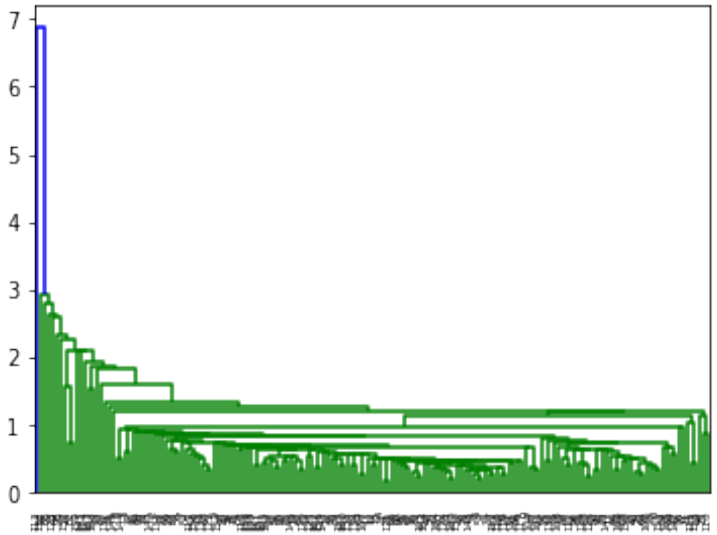


Task 7: Hierarchical Clustering

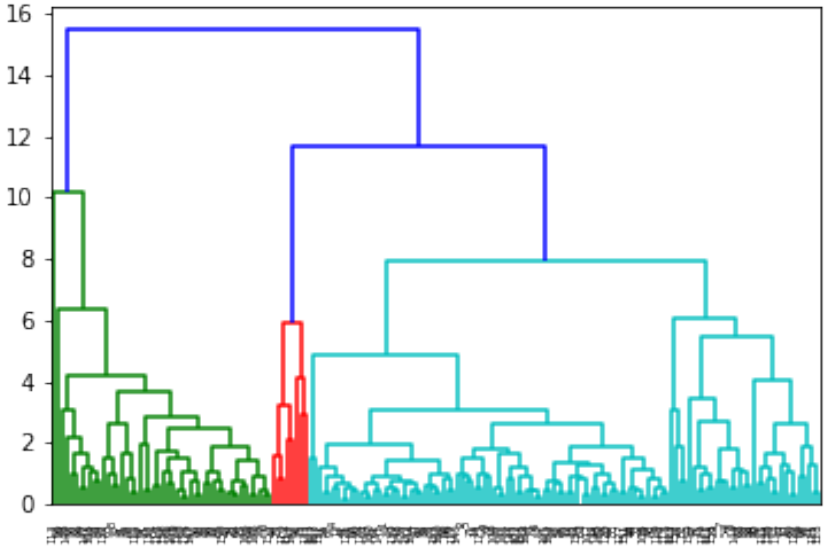
Single Linkage : The following are observed from the above clustering Profile and Value counts of Single Linkage : Only one cluster is dominating other clusters i.e. cluster label 0
The total count of cluster label 0 is 165
other cluster label is having one count each
Hence Single Linkage Cluster is not used in the further anal**ysis**

Complete Linkage:Similar points can be observed as seen in KMeans Clusters:
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

Single Linkage



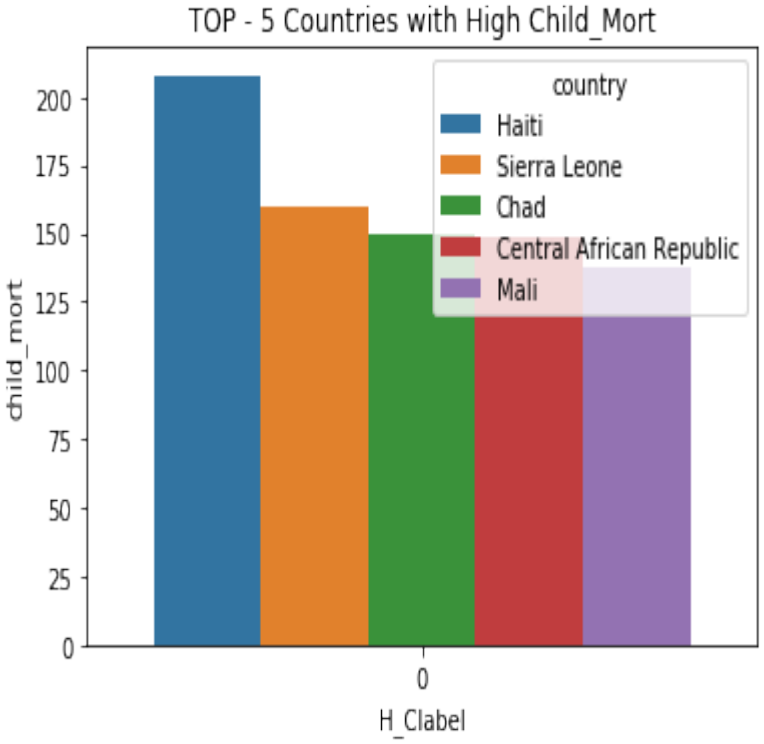
Complete Linkage:



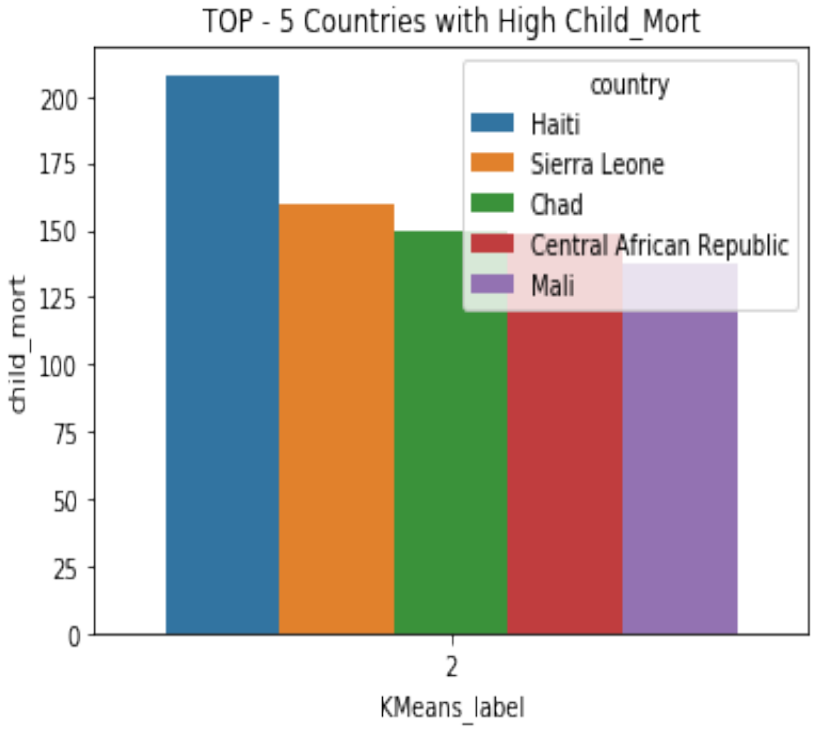
Task 8:

Comparing the K Means and Hierarchical Clustering

plotting top 5 countries where child_mort is high using Hierarchical Cluster (Complete Linkage



plotting top 5 countries where child_mort is high using KMeans Clustering



From the above comparison, Both clusters show the same top - 5 countries. They are:(from highest child_mort to least)
Haiti
Sierra Leone
Chad
Central African Republic
Mali

Conclusion :

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
The top 5 Countries which are in need of direst AID are:
Haiti
Sierra Leone
Chad
Central African Republic
Mali
The most of African countries are in top - 5 which are in need of direst AID

THANK
YOU