Help International

CLUSTERING ASSIGNMENT

AUTHOR:- SRINIVAS AVALA

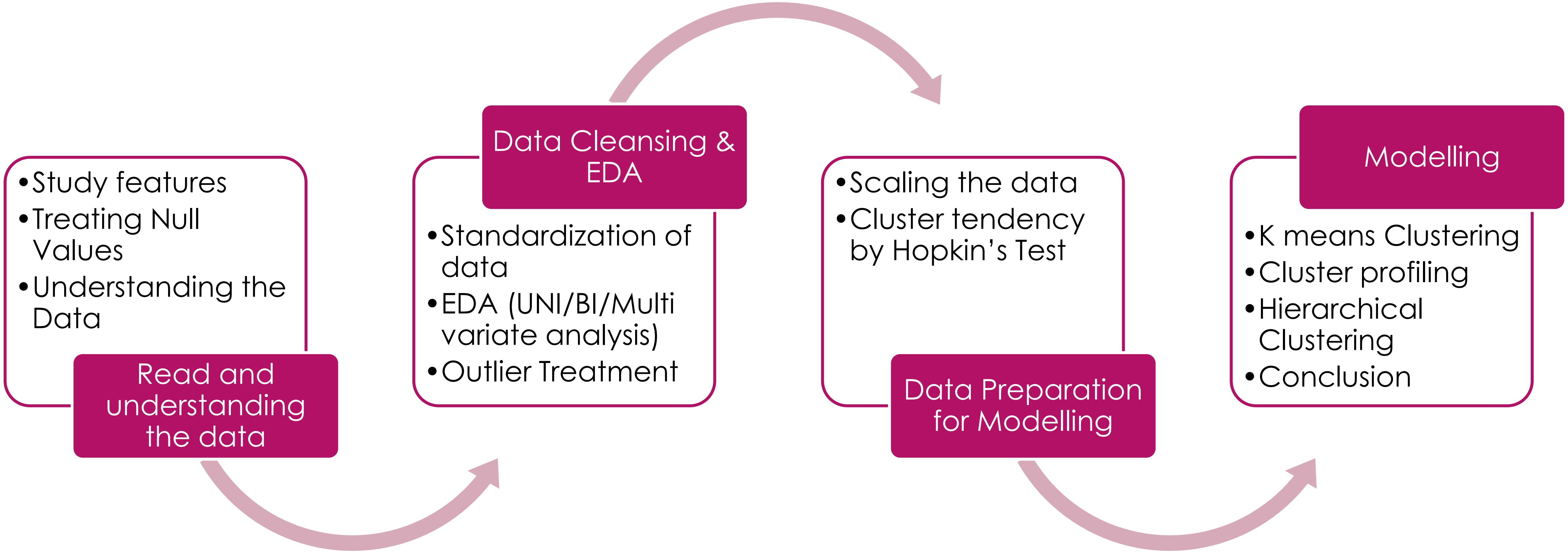
Objective

- HELP international is an international humanitarian NGO that is providing the people of backward countries with basic needs, amenities and relief during the time of disasters and natural calamities.
- HELP has collected the data of various countries that explains the socio-economic and health factors of them , based on which they would like to shortlist the top 5 countries that needs the Aid.

Features

- child_mort
- exports
- health
- imports
- income
- inflation
- life_expec
- total_fer
- gdpp
- country

End to End Flow



Read and understanding the data

- Study features
- Treating Null Values
- Understanding the Data

Data Cleansing & EDA

- Standardization of data
- EDA (UNI/BI/Multi variate analysis)
- Outlier Treatment

Data Preparation for Modelling

- Scaling the data
- Cluster tendency by Hopkin's Test

Modelling

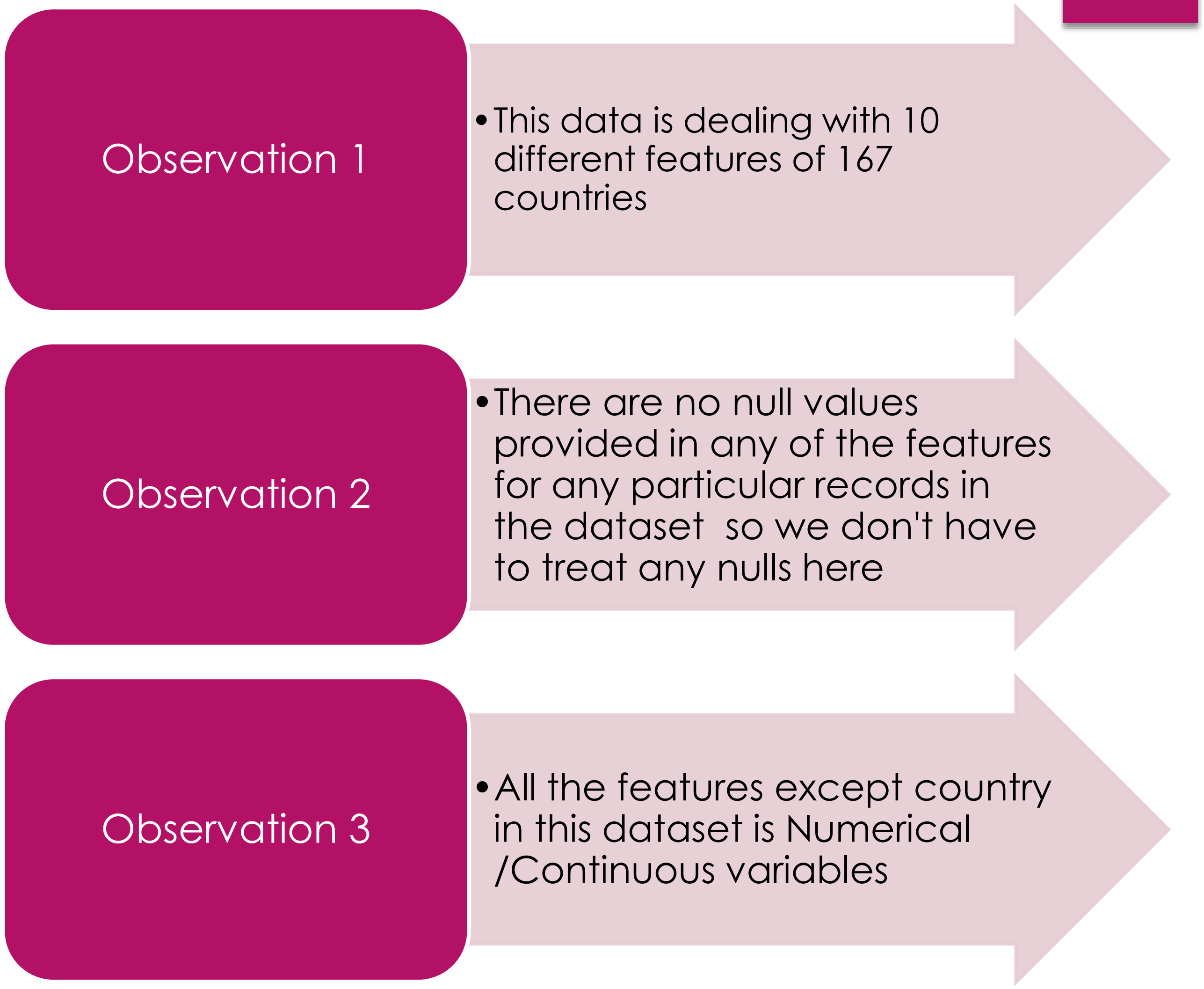
- K means Clustering
- Cluster profiling
- Hierarchical Clustering
- Conclusion

Observations/ Task Performed

Observations of the data based on the analysis done by the Data scientist in various stages

This slide talks about the first/initial step of analysis which is read and understand the data

Read and Understand the data



Observation 1

- This data is dealing with 10 different features of 167 countries

Observation 2

- There are no null values provided in any of the features for any particular records in the dataset so we don't have to treat any nulls here

Observation 3

- All the features except country in this dataset is Numerical /Continuous variables

Observations/ Task Performed

Observations of the Data based on the analysis done by the Data scientist in various stages

This slide talks about the second step of analysis of the data which is Data Cleansing & EDA

Data Cleansing & EDA

Standardization of Data

- Identified that there are few features like exports , health ,imports are provided as the percentage values of GDPP , so converted them as standard values rather than using as percentage values. This helps during scaling the data

EDA(Uni/Bi/Multi variate analysis)

- Performed EDA on entire dataset to understand the nature of data and implemented the data Visualization by Uni/Bi/multi variate analysis and Correlation between the features
- Observed that features like child_mort, health, inflation are normally distributed
- Gdpp has strong correlation between income, import ,export and health
- Imports and Exports has strong correlation along with child rate and total fertility

Data Cleansing and EDA

Observations/ Task Performed

Observations of the Data based on the analysis done by the Data scientist in various stages

This slide talks about the second step of analysis of the data which is Data Cleansing & EDA

Outlier Treatment

- Plotted the Box plot for all the features and identified that there are Outlier in all the Variables
- Observed that Outliers are available on the upper fence for all the features except for life_expec
- Capped the features available in the Upper fence to 0.99 and 0.98 quantile so that most of the
- Capped life_expec to 0.05 value so that it is aligned with the remaining values

Data preparation for Modelling

Observations/ Task Performed

Observations of the Data based on the analysis done by the Data scientist in various stages

This slide talks about the third step of analysis of the data which is Data Preparation for Modelling

Scaling the data:-

Scaled the data using standard scalar function and made sure that standard deviation and mean values are normalized

Hopkins's Test:-

Performed the Hopkin's test on the normalized data and identified the cluster tendency is around 0.85 which is good.

Modelling (K means)

Observations/ Task Performed

Observations of the data based on the analysis done by the Data scientist in various stages

This slide talks about the fourth step of analysis of the data which is Data Preparation for Modelling

Identifying K value -

- Plotted Silhouette score plot and Elbow Curve and Observed that the most silhouette score achieved was 0.47
- Based on both methods we identified that the optimal K value for this process is 3

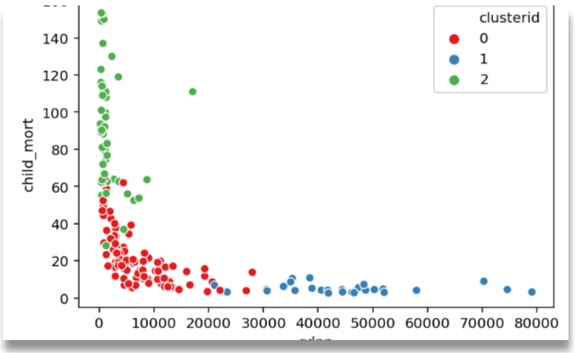
K means Clustering:-

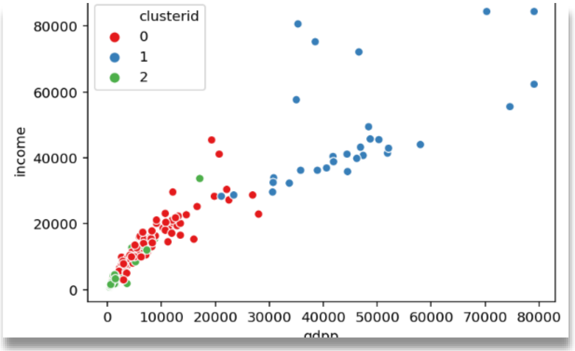
- Identified the labels by using the scaled data and assigned the labels to the Countries from the original dataset.
- Plotted the data points between 2 set of main features and observed that the cluster formation is proper.

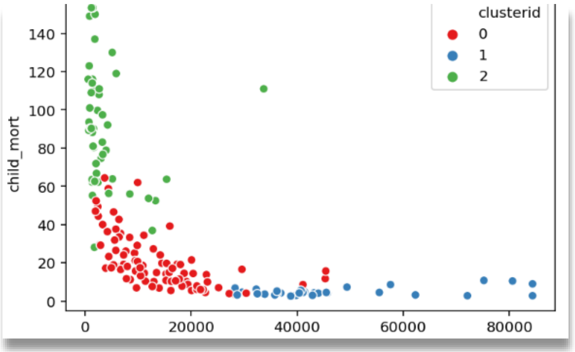
Cluster Profiling:-

- Plotted and Observed the mean values of each features per cluster
- Plotted separate bar chart with main features for cluster selection using income, gdpp, child_mort as required and identified that Cluster 2 have the less mean value , We have considered that set for our next steps as we need to identify the countries that needs aid

K Means Clustering for important features







Child_mort vs gdpp

The plot between child_mort and gdpp is as shown above and the cluster 0 ,1 and 2 have slight integration but looks like the segregation happened properly

Income vs gdpp

The plot between income and gdpp is as shown above and the cluster 0 ,2 have significant amount of integration where as cluster 1 has complete independent cluster

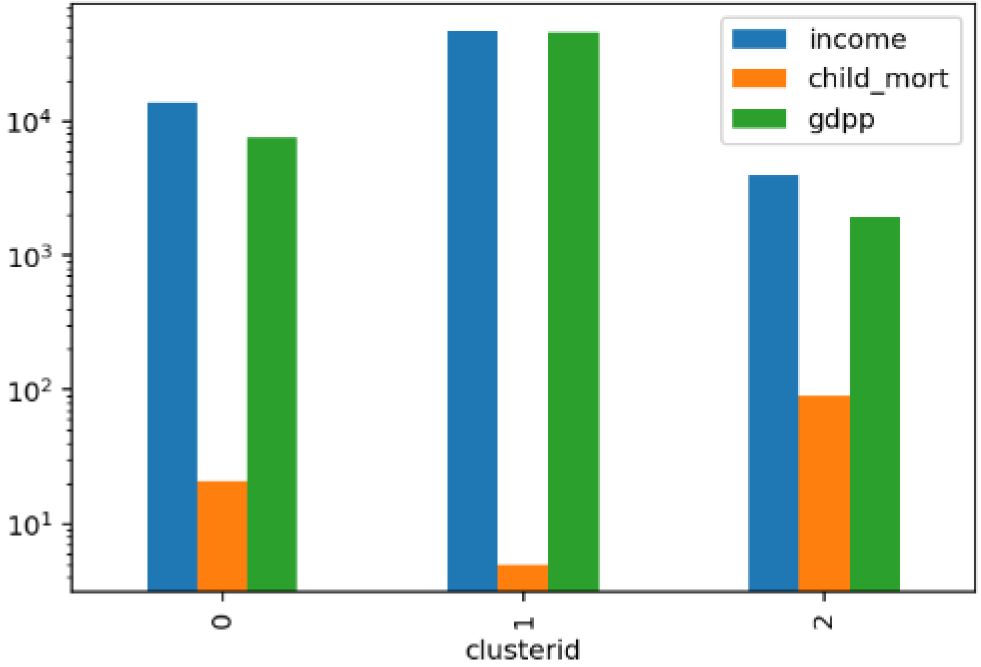
Child_mort vs income

The plot between child_mort and income is as shown above and the cluster 0 ,1 and 2 have slight integration but looks like the segregation happened properly

Income vs child_mort vs gdpp obtained by K- means

Plotted the graph between income, child_mort and gdpp , have considered the logarithmic scale since the values of child_mort is very less compared to other features.

Observed that cluster 2 has less income and gdpp values with more child_mort which we need to consider as per of business case so the cluster we need to opt is 2.



Modelling(Hierarchical)

Observations/ Task Performed

Observations of the Data based on the analysis done by the Data scientist in various stages

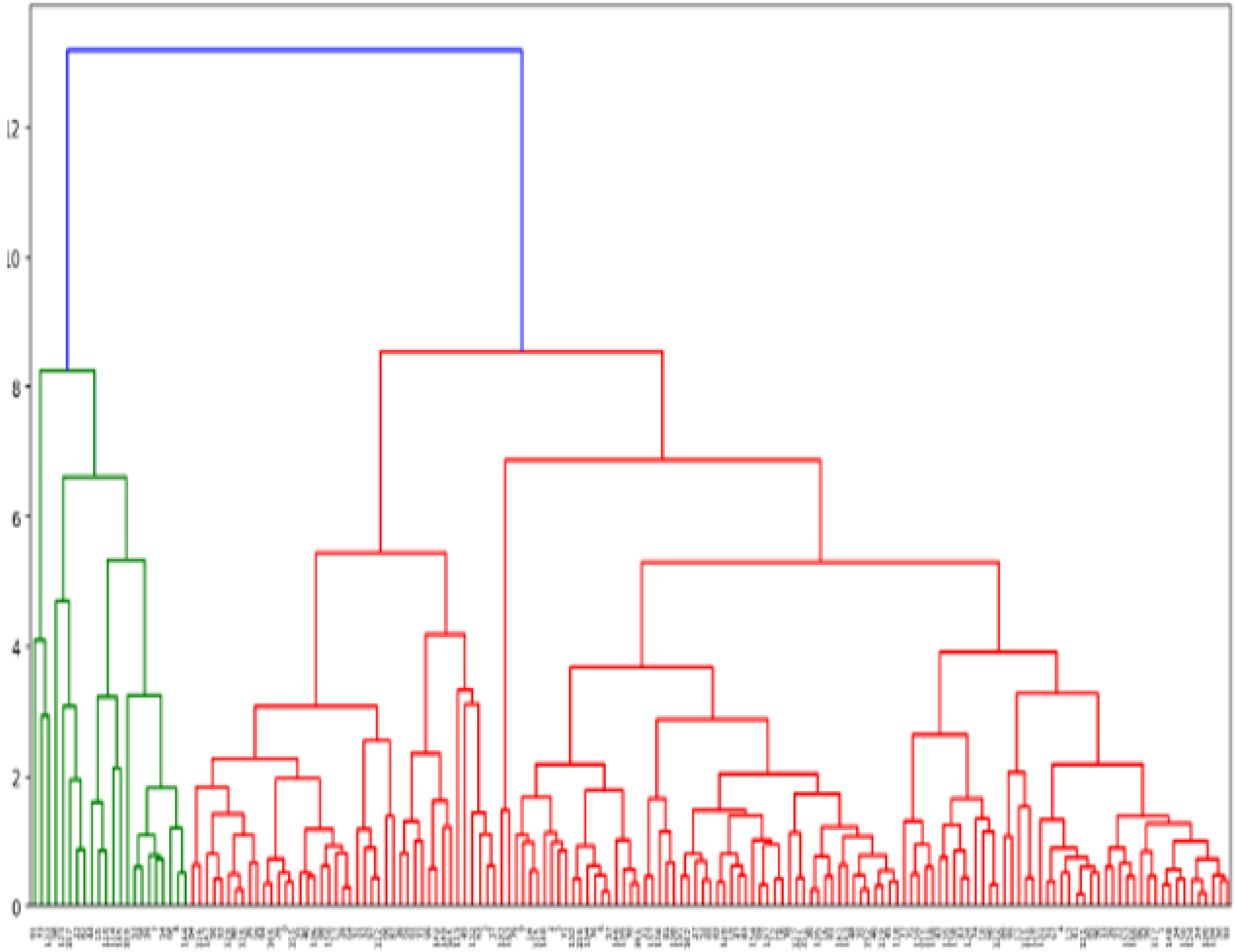
This slide talks about the fourth step of analysis of the data which is Data Preparation for Modelling

Hierarchical Clustering-

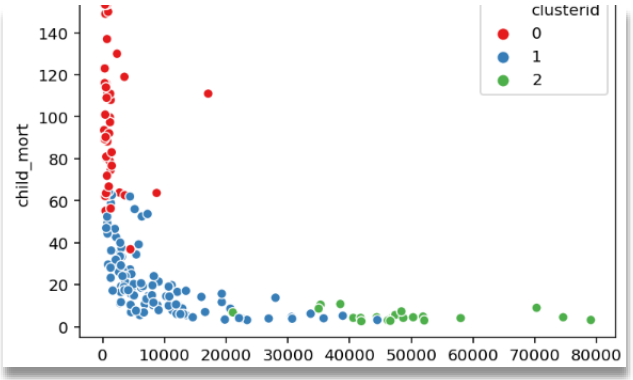
- Implemented single linkage and complete linkage dendrogram for the scaled dataset by choosing the number of clusters as 3.
- Created a new dataset by merging the cluster labels achieved from the dendrogram
- Observed that the mean value for cluster labelled 0 is having less values so considered cluster 0 for the next steps
- Plotted the scatter plot using the important features and made sure that the clusters are formed properly

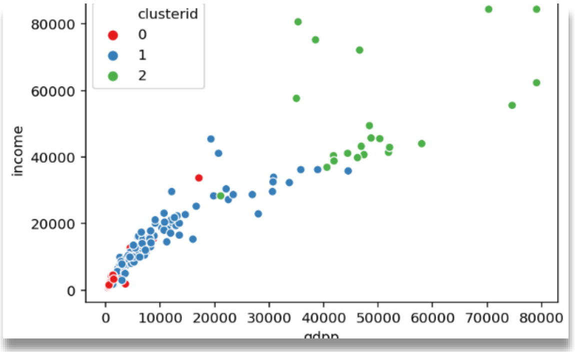
Dendrogram

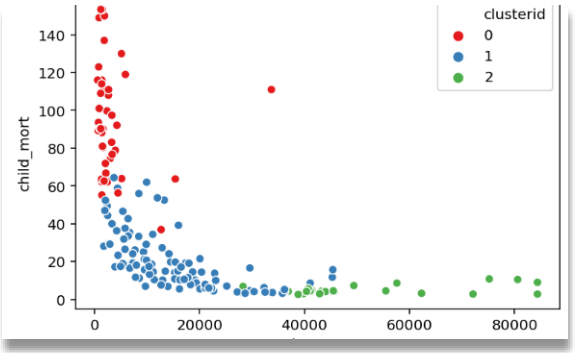
The dendrogram explains the various levels of cluster formation using hierarchical Clustering mechanism



Hierarchical Clustering for important features







Child_mort vs gdpp

The plot between child_mort and gdpp is as shown above and the cluster 0 ,1 and 2 have slight integration but looks like the segregation happened properly

Income vs gdpp

The plot between income and gdpp is as shown above and the cluster 0 ,2 have significant amount of integration where as cluster 1 has complete independent cluster

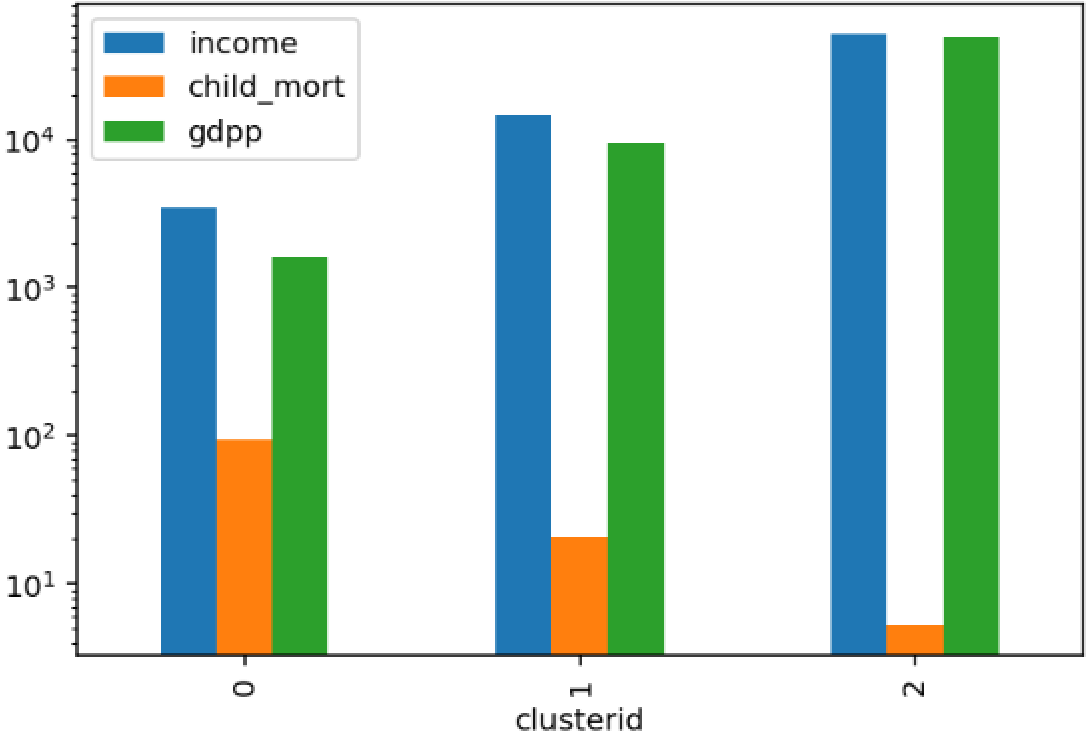
Child_mort vs income

The plot between child_mort and income is as shown above and the cluster 0 ,1 and 2 have slight integration but looks like the segregation happened properly

Income vs child_mort vs gdpp obtained by Hierarchical

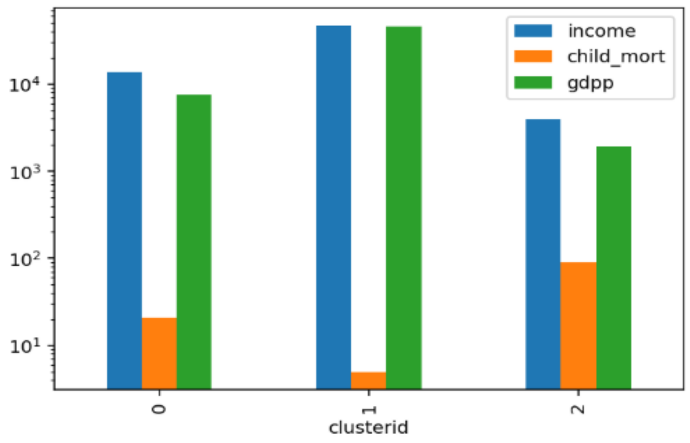
Plotted the graph between income, child_mort and gdpp , have considered the logarithmic scale since the values of child_mort is very less compared to other features.

Observed that cluster 0 has less income and gdpp values with more child_mort as per hierarchical clustering approach which Business would like to consider so the cluster we need to opt is 0.

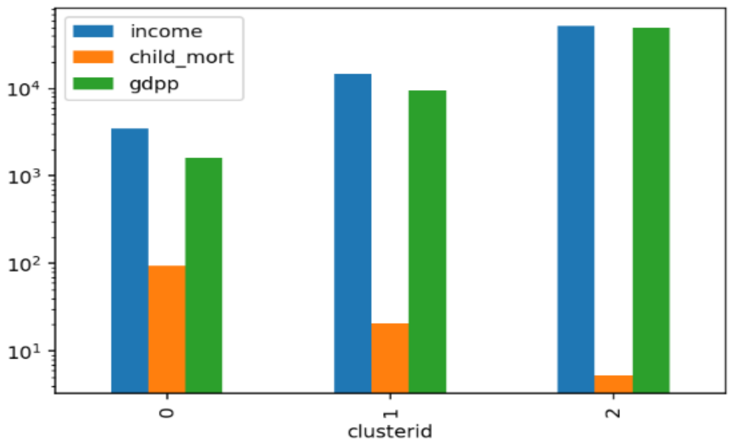


K means vs Hierarchical

K means



Hierarchical



The clusters labels for K means may be different from the clusters formed in Hierarchical , But the mean ranges of the clusters are identical.

Conclusion:-

As per the business requirement we need to identify the countries that required AID from HELP international , so based on the K means Clustering and Hierarchical clustering, I've identified the top 5 countries as below.

country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
Burundi	93.6	20.6052	26.7960	90.552	764.0	12.30	57.70	6.2600	231.0
Liberia	89.3	62.4570	38.5860	302.802	700.0	5.47	60.80	5.0200	327.0
Congo, Dem. Rep.	116.0	137.2740	26.4194	165.664	609.0	20.80	57.50	6.5400	334.0
Niger	123.0	77.2560	17.9568	170.868	814.0	2.55	58.80	6.5636	348.0
Sierra Leone	153.4	67.0320	52.2690	137.655	1220.0	17.20	55.78	5.2000	399.0

Conclusion:-

- By seeing the data in the previous table we can understand that the Gdpp value for all these countries are very less compared to other countries
- They are exporting less and Importing more which explains that these countries do not have proper daily requirements
- These countries do not have enough funds to spend for health compared to other countries

I conclude that these countries are the most eligible to get AID from HELP International

Russia
Norway
United Kingdom
Poland
Germany
Ukraine
France
Italy
Spain
Kazakhstan
Mongolia
North Atlantic Ocean
Turkey
China
Japan
South Korea
Iraq
Iran
Afghanistan
Algeria
Libya
Egypt
Pakistan
Saudi Arabia
India
Mali
Niger
Thailand
Sudan
Chad
Nigeria
Ethiopia
Venezuela
mbia
Kenya
DRC
Indonesia
Papua New Guinea
Tanzania
Brazil
ru
Angola
Bolivia
Namibia
Indian

Thank You…!!!!!