Unsupervised Clustering Assignment

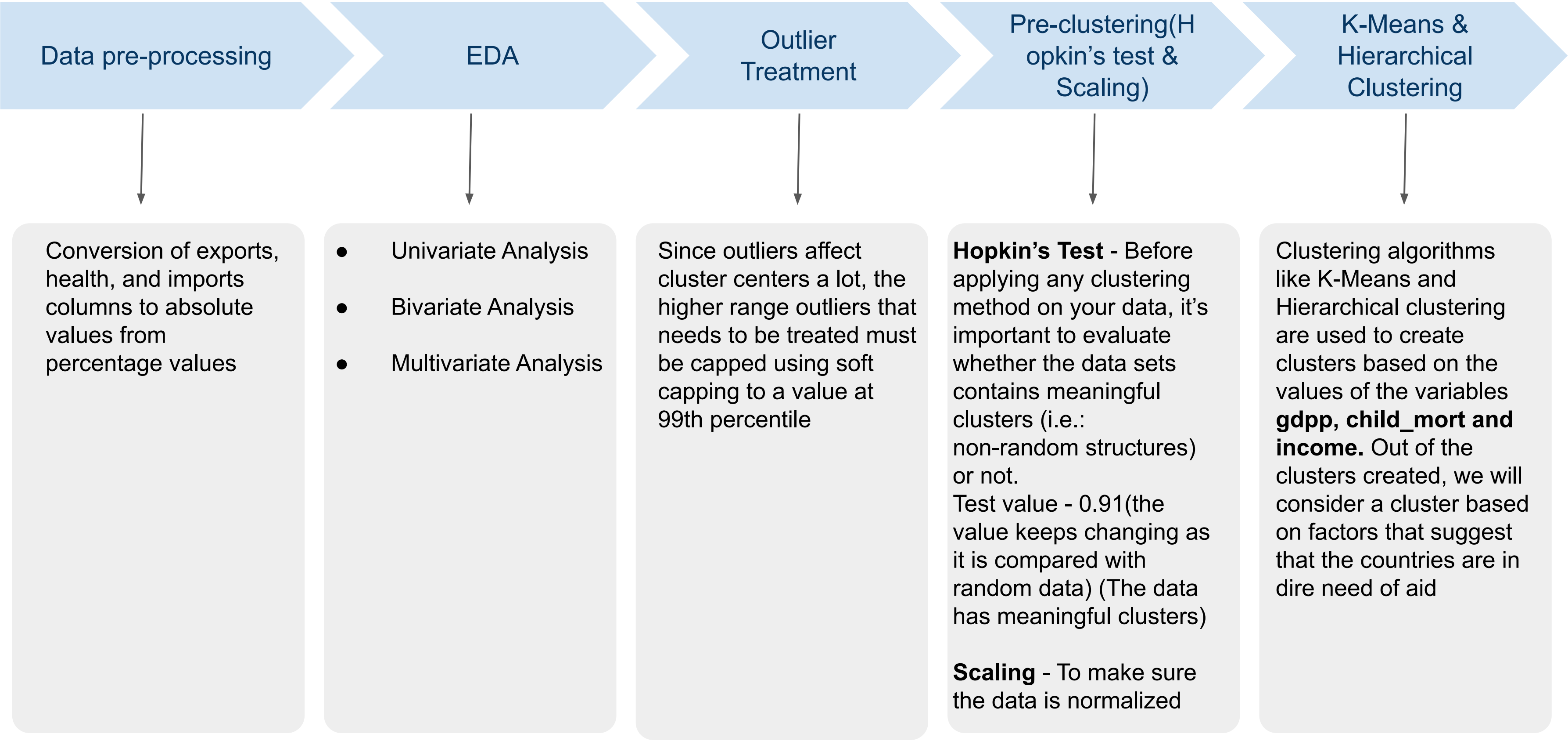
Name:- Divit Karmiani
Email ID:- divitkarmiani1998@gmail.com

Problem Statement

The NGO, **HELP International - an international humanitarian NGO**, needs to decide how to use the funds of **$10 million 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**.

Task - To suggest the countries which are in dire need of aid, which the CEO needs to focus, by **categorising** the countries using some socio-economic and health factors that determine the overall development of the country.

Analysis Approach



Data pre-processing

Conversion of exports, health, and imports columns to absolute values from percentage values

EDA

- Univariate Analysis
- Bivariate Analysis
- Multivariate Analysis

Outlier Treatment

Since outliers affect cluster centers a lot, the higher range outliers that needs to be treated must be capped using soft capping to a value at 99th percentile

Pre-clustering(Hopkin's test & Scaling)

Hopkin's Test - Before applying any clustering method on your data, it's important to evaluate whether the data sets contains meaningful clusters (i.e.: non-random structures) or not.
Test value - 0.91(the value keeps changing as it is compared with random data) (The data has meaningful clusters)

Scaling - To make sure the data is normalized

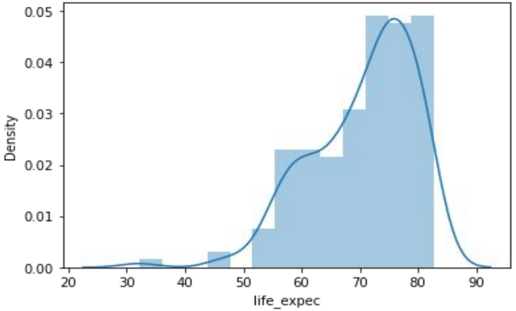
K-Means & Hierarchical Clustering

Clustering algorithms like K-Means and Hierarchical clustering are used to create clusters based on the values of the variables **gdpp, child_mort and income.** Out of the clusters created, we will consider a cluster based on factors that suggest that the countries are in dire need of aid

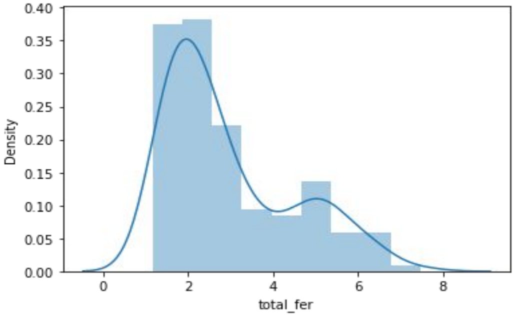
Exploratory Data Analysis

Univariate

Life Expectancy Histogram

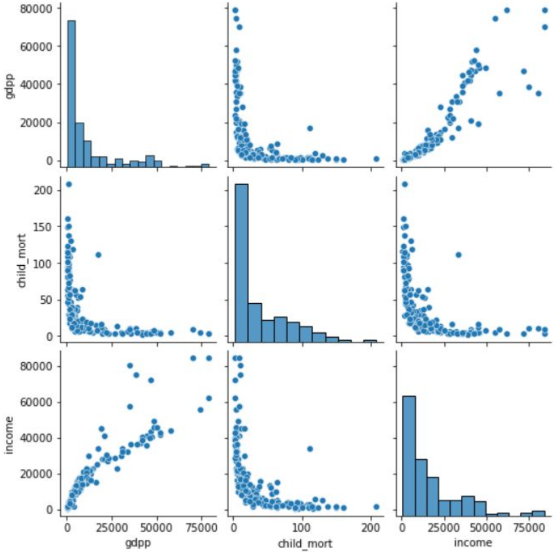


Total Fertility Histogram



Showcases internal groupings in each feature

Bivariate

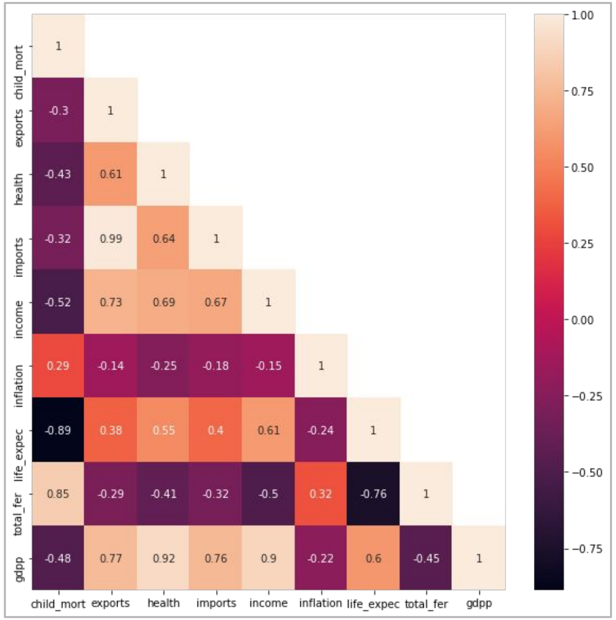


Pairplot for features gdpp,child_mort,income

Similarly pairplot of other features were plotted.

Showcases relationship between features taken 2 at a time

Multivariate



Showcases correlations between features

Exploratory Data Analysis - Learnings

Univariate

- Most of the variables follow normal distribution with no internal groupings.

- Except for variables like income,gdpp,total_fer and life_expec. The distributions indicate some internal groupings and hence indicate some cluster formation.

Bivariate

- gdpp almost follows a linear relationship with variables like exports, health, imports, income. This is obvious in terms that as exports, imports, income increase, gdpp also increases

- We also see that for higher gdpp, the values of life_expec are very high.

- While for higher gdpp, we see values of child_mort and total_fer are very low.

- As expenditure on health increases, the life_expec also tends to be higher. This is obvious as well.

Multivariate

- We see high correlations of variables exports, health, imports, income with gdpp. This is in line with what we saw in bivariate analysis.
- We also see high positive correlations between
  - exports and imports which is obvious.
  - exports,health and imports with income
- We see negative correlation between
  - child_mort and life_expec :- This is obvious. As child_mortality rate increases, life_expectancy decreases
  - total_fer and life_expec :- This is good information to derive insights from. Maybe as the total_fer increases, it results in population boom, which might be responsible for lowering of life_expec. Maybe because of shortage of resources for each individual.

Outlier Treatment

For columns child_mort, inflation and total_fer, we should not treat the higher range outliers as it is critical to our business use case. These columns, if the values are high, suggest that the countries are in dire need of aid.

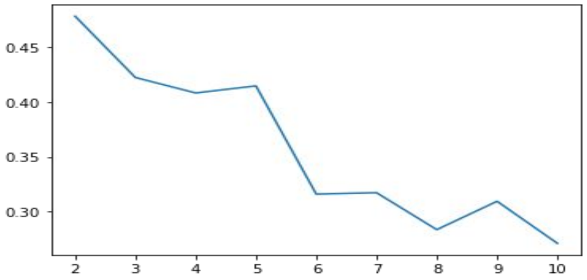
For other columns, we should not treat the lower range outliers for the same reason of them being critical to our business use case. These columns, if the values are low, suggest that the countries are in dire need of aid.

Considering above points, outlier treatment was done.

K-Means Clustering -1

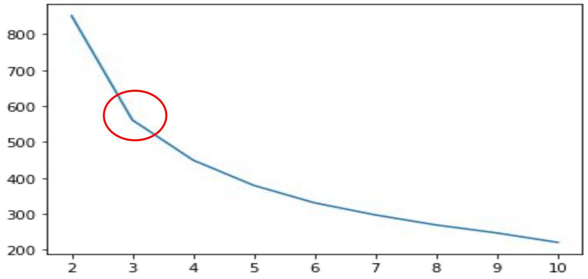
Deciding the number of clusters (k)

Silhouette score



The Silhouette score from 2 to 10 are plotted in the graph above. We will consider the value of k with highest silhouette score (k=2), and therefore we take the next best, i.e. k=3

SSD



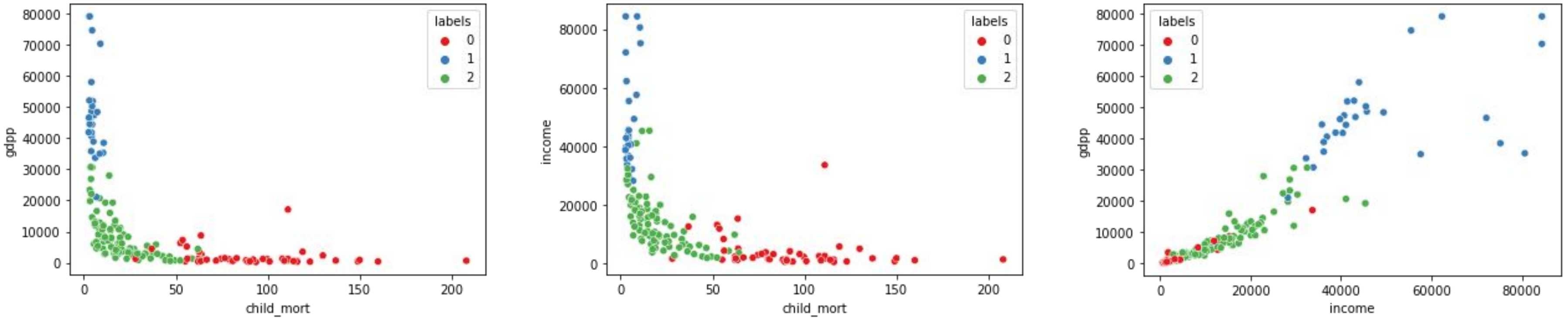
In this case, the elbow curve is plotted and we will consider the value of k at the bend of the elbow (k=3)

Final value of k= 3 +- 1. k=2 clusters doesn't make much sense. Hence we will only go with k=3 and k=4 clusters.

K-Means Clustering -2

Visualising and analysing clusters using scatter plots for variables gdpp, child_mort, income

K = 3

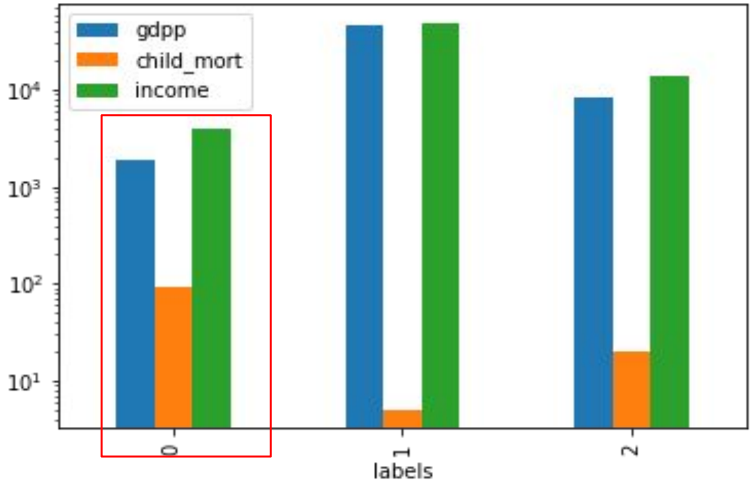


Since for k=4 clusters, silhouette score is less than k=3 clusters and while visualising the clusters as well, I found that for k=3 the information captured is clearly distinguishable and is able to capture information in 3 clusters only, there is no need for a 4th cluster. We will go ahead with 3 clusters only.

K-Means Clustering -3

Cluster profiling and identifying the correct cluster

Cluster Profiling



The cluster with label 0 is the ideal cluster for our business use case. It has lowest gdpp, highest child_mort, and lowest income in average amongst all other clusters. The countries in this cluster are the ones in dire need of aid.

Final countries using K-Means

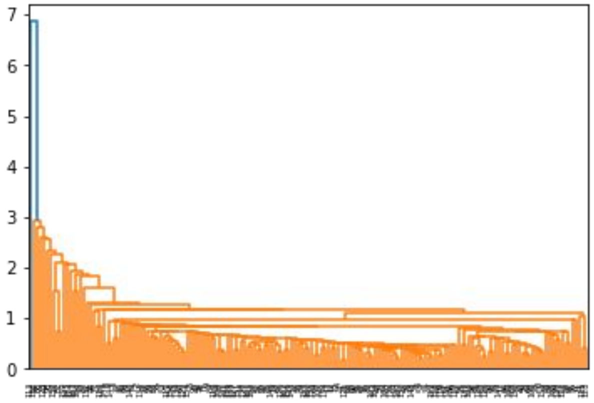
Top 5 countries in dire need of aid on the basis of gdpp, child_mort and income(using k-means clustering) are:

1. Burundi
2. Liberia
3. Congo, Dem. Rep.
4. Niger
5. Sierra Leone

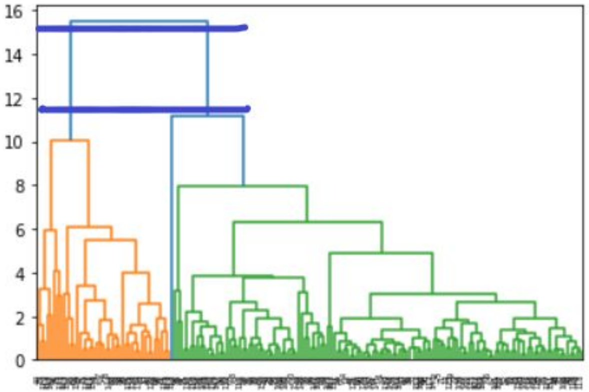
Hierarchical Clustering -1

Deciding the number of clusters (k)

Single Linkage



Complete Linkage

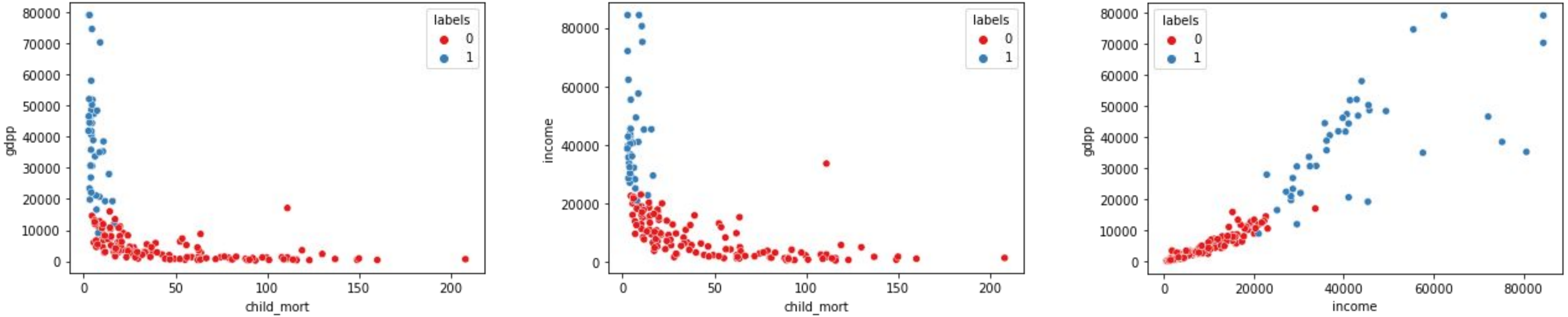


- **Complete linkage gives better dendrogram than single linkage.**
- **If we see the cluster labels for k=3 and k=4, we see only 1 point in the last cluster. One point cannot constitute a cluster.**
- **Also if we interpret the dendrogram, for k=2, it has the maximum vertical distance as shown in below figure as compared to k=3 or k=4**

Hierarchical Clustering -2

Visualising and analysing clusters using scatter plots for variables gdpp, child_mort, income

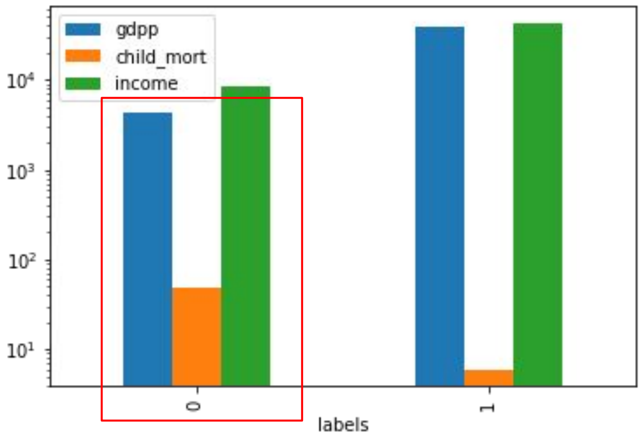
K = 2



Hierarchical Clustering -3

Cluster profiling and identifying the correct cluster

Cluster Profiling



The cluster with label 0 is the ideal cluster for our business use case. It has lowest gdpp, highest child_mort, and lowest income in average amongst all other clusters. The countries in this cluster are the ones in dire need of aid.

Final countries using K-Means

Top 5 countries in dire need of aid on the basis of gdpp, child_mort and income(using hierarchical clustering) are

1. Burundi
2. Liberia
3. Congo, Dem. Rep.
4. Niger
5. Sierra Leone

Final Results

- We have used clustering methods - K-Means and Hierarchical to categorize the countries based on the variables **gdpp, child_mort and income**

- Using K-means we categorized the countries in 3 clusters and using Hierarchical clustering, we categorized the countries in 2 clusters, both giving the same final results

- Based on the clustering, we have found that the countries that fall in the clusters with lowest gdpp, highest child_mort, and lowest income are the ones that are in direst need of aid and the funds should be allocated for their help

- The countries identified are as follows:
  a. Burundi
  b. Liberia
  c. Congo, Dem. Rep.
  d. Niger
  e. Sierra Leone