Clustering Algorithm

Present By : Subhasis Pattanayak

Business case

- HELP International is an international humanitarian NGO that is committed to fighting poverty and providing the people of backward countries with basic amenities and relief during the time of disasters and natural calamities. It runs a lot of operational projects from time to time along with advocacy drives to raise awareness as well as for funding purposes.

Problem Statement

- After the recent funding programmer, they have been able to raise around $ 10 million. Now the CEO of the NGO needs to decide how to use this money 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.

Project Goal & Objectives

- To categories the countries using some socio-economic and health factors that determine the overall development of the country.
- To suggest the countries which the CEO needs to focus on the most.

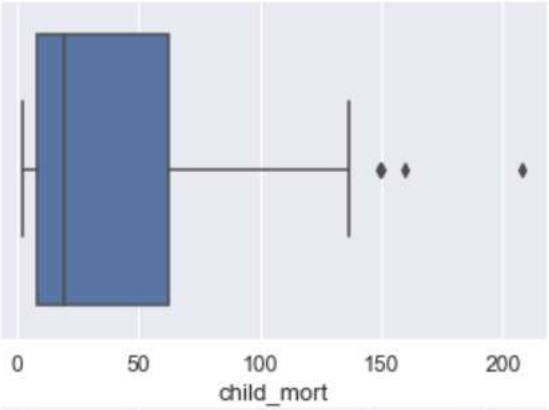
Technical Steps

- Basic Sanity checks
- Exploratory Data Analysis
  - Outliers Treatment
  - Univariate Analysis
  - Bivariate Analysis
- Scaling
- Hopkins Statistics
- Elbow Curve / SSD
- Silhouette Analysis
- Profiling using K-Means labels with respect to three variables (Child_mort, gdpp, income)
- Hierarchical Clustering
  - Single Linkage & Complete Linkage
  - Finding the Clusters using Cut-tree method
  - Profiling using Hierarchical labels with respect to three variables (Child_mort, gdpp, income)
- Conclusion & Recommendation

Exploratory Data Analysis

Box plot for Outliers detection :-

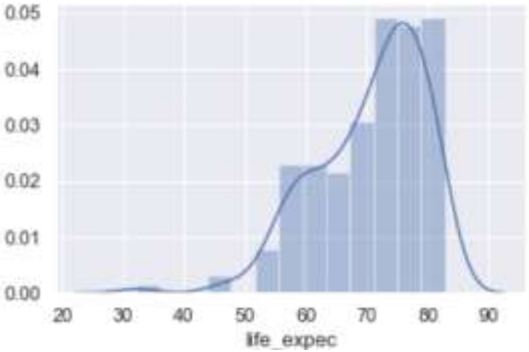
- As the data set contains (167, 10) small amount of data better to cap the outliers.
- We treated the lower end values of the child_mort variables because if we cap the upper end outliers values which are the countries that are in the direst need of aid.
- rest of all the outliers of the variables except child_mort we capped that of upper end.



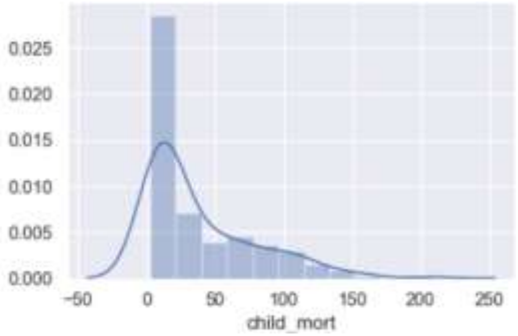
EDA – Univariate Analysis

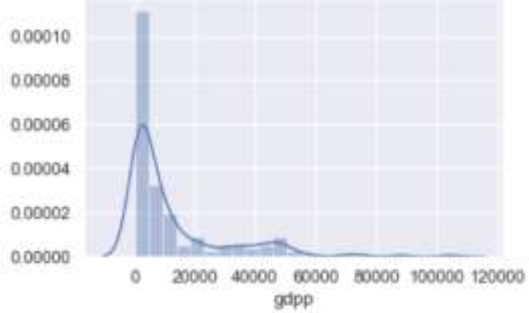
In Univariate Analysis we try to find the distribution using dist plot.

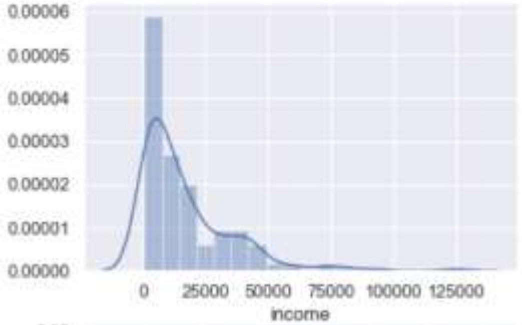
- Most the variables are not in normally distribution.
- Most of the variable are positively skewed.
- Only life_expec shows the negatively skewed.



Negatively skewed





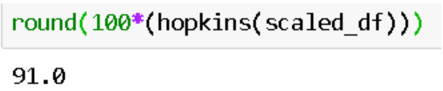


Hopkins statistics, Elbow curve and Silhouette Analysis

- Used Hopkins Statistics to test the dataset is suitable for clustering or not

Applied the Hopkins Statistics on the scaled data frame.

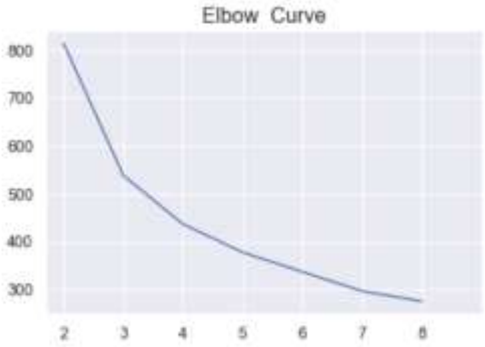
- We got 88 to 91 % which is the good score for clustering.



```
round(100*(hopkins(scaled_df)))
```

```
91.0
```

- In Elbow curve we got 3 or 4 is good for the value of K.

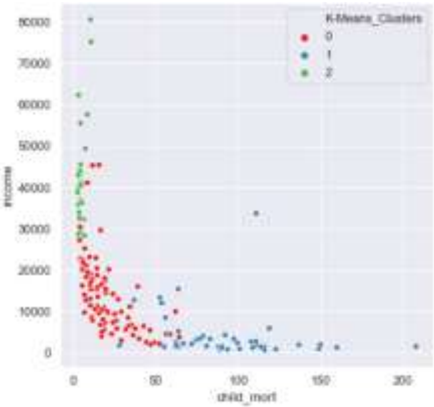


- Silhouette score analysis to find the ideal number of clusters for K-means clustering.
- we got 3 is the ideal number for Clustering.

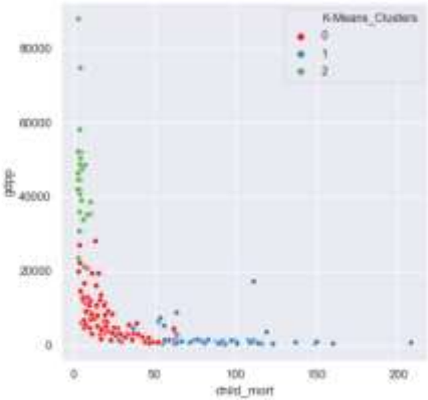
n_Cluster	Silhouette Score
2	48
3	40
4	39
5	38
6	29
7	29
8	32

K-Means Clusters

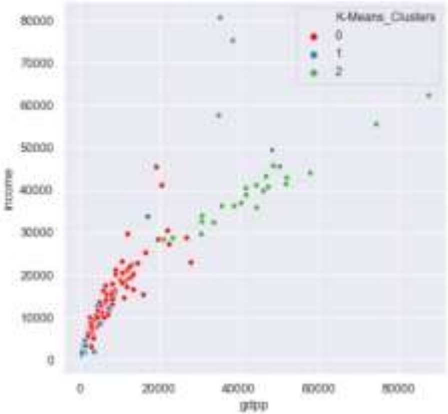
- Using the Scatter plot we distinguish the Cluster labels as per the K-Means algorithm with the three important variable which are (child_mort, gdpp, income)
- **Cluster labels**
- Cluster 0 – 89 Cluster 1 - 47 Cluster 2 - 27



Plot 1 – child_mort vs. income



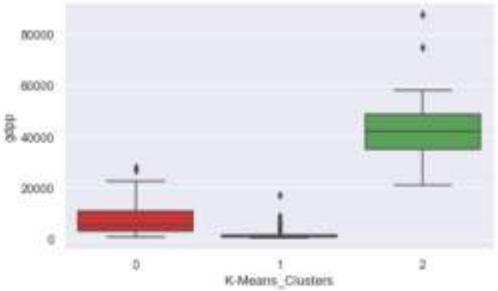
Plot 2 – child_mort vs. gdpp

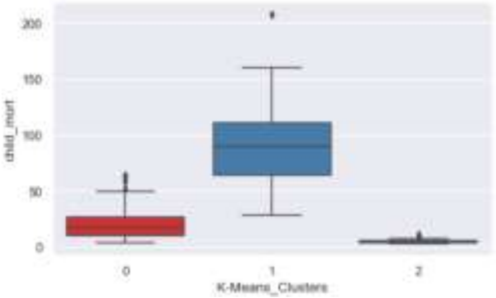


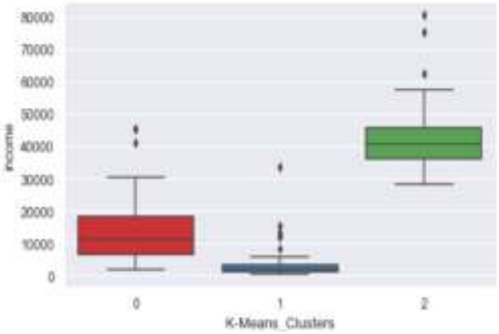
Plot 3 – gdpp vs. income

Cluster Profiling(K-Means)

- by using K-Means Clustering we profiled the cluster with respect to 3 variables (child_mort, gdpp, income) using Box plot.
- Cluster 1 represents the high child_mort low income, low gdpp.
- Cluster 2 represents High income, high gdpp and low child_mort
- Cluster 0 represents moderate in all cases like child_mort, income, gdpp

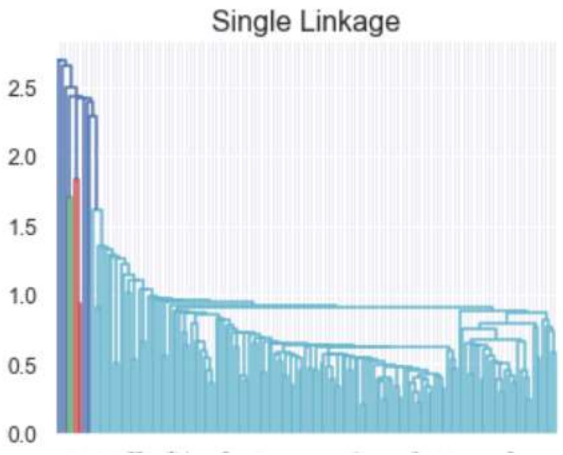


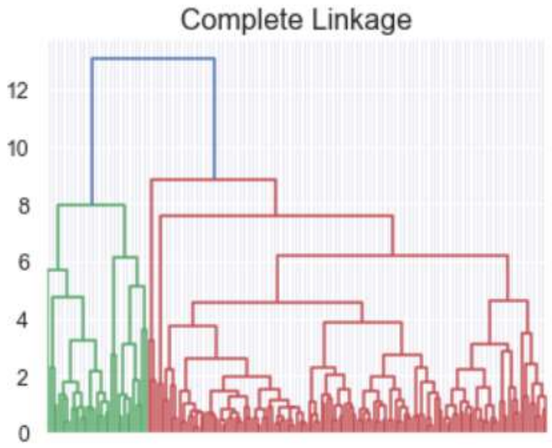




Hierarchical Clustering

- This shows how many Clusters the data can be split into.

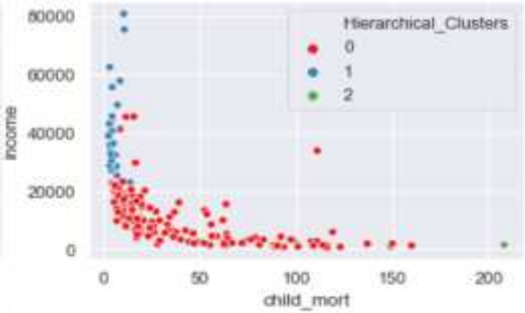




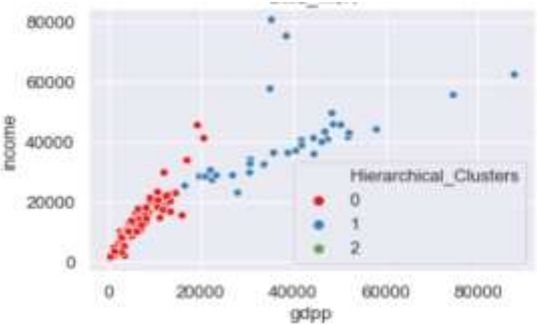
- In Single linkage it takes two closest points of two clusters and defined that as the distance.
- In Complete Linkage it’s showing perfectly to choose how many clusters can create.
- Here Complete Linkage showing better than Single Linkage.

Hierarchical Clusters

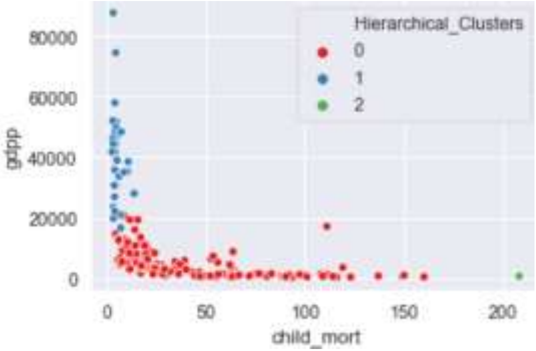
- Using the Scatter plot we distinguish the Cluster labels as per the Hierarchical algorithm with the three important variable which are (child_mort, gdpp, income)
- **Cluster labels(Number of Clusters)**
- Cluster 0 – 127 Cluster 1 - 33 Cluster 2 - 3



Plot 1 – child_mort vs. income



Plot 2 – gdpp vs. income



Plot 2 – child_mort vs. gdpp

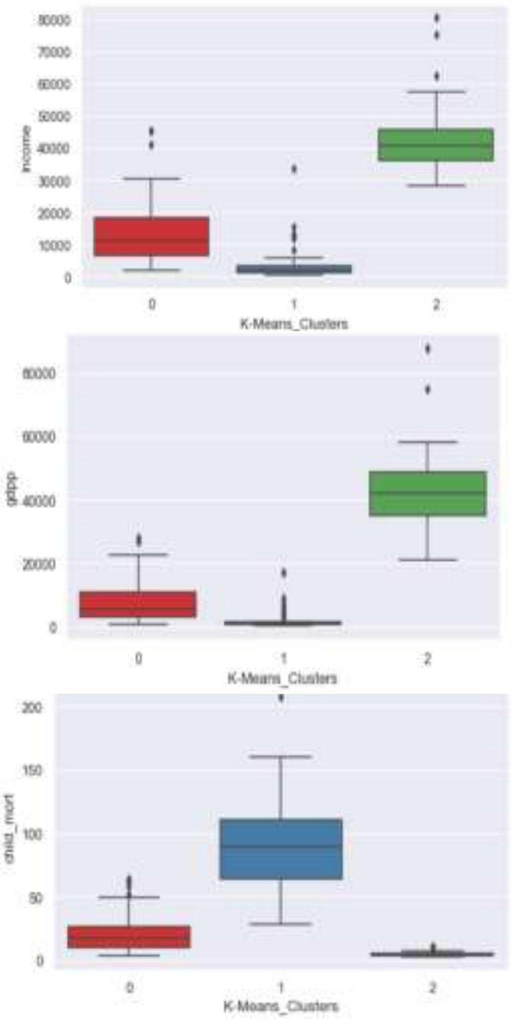
Cluster Profiling(Hierarchical)

- by using Hierarchical Clustering we profiled the cluster with respect to 3 variables (child_mort, gdpp, income) using Box plot.

- Cluster 1 represents the high child_mort low income, low gdpp.

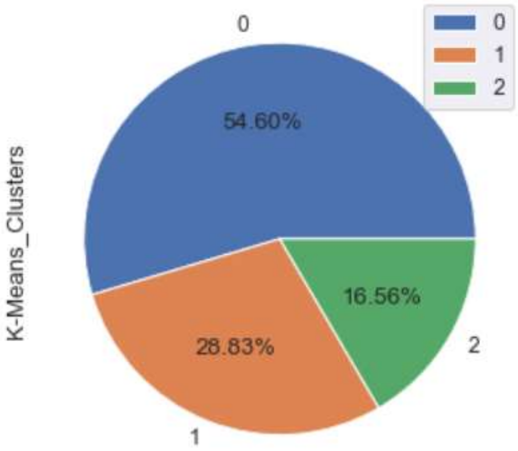
- Cluster 2 represents High income, high gdpp and low child_mort

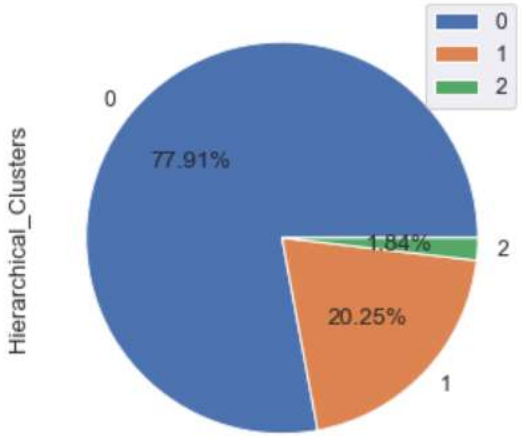
- Cluster 0 represents moderate in all cases like child_mort, income, gdpp



K-Means vs. Hierarchical

- In both of the Clustering algorithmm Cluster 0 is high
- Medium is Cluster 1
- Cluster 2 is low but very less in Hierarchical algorithm





Conclusion & Recommendations

- The top 5 countries that require help the most are listed below :
  1 - Burundi, 2 - Liberia, 3 - Congo, Dem. Rep, 4 - Niger 5 - Sierra Leone
- These countries have
  - very low rate of income per person, gdpp per capita, average number of a new born child would live.
  - very high rate of measurement of the annual growth rate, number of children that would be born and child mortality
  - It is clear that the NGO sholud work above countries need very quick aid in terms of money, education and services.

☺ **THANKS!**