



Clustering Assignment-Part I

K-Means + Hierarchical

Submitted By : Snigdha Chakraborty




Background

- 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
- After the recent funding programs, 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



Problem Statement

- To categorize 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
 - To report back at least 5 countries which are in direst need of aid from the analysis work that we performed
- 



Technical Approach

- Data Understanding and Data Cleaning
- EDA
 - Univariate Analysis
 - Bivariate Analysis
- Data Preparation
 - Outlier Treatment
 - Data Scaling
 - Hopkins Statistics (to check if the dataset is good enough for a cluster analysis)
- Clustering
 - K-Means Clustering
 - Silhouette Score
 - Elbow Curve/SSD
 - Hierarchical clustering
 - Single Linkage
 - Complete Linkage
- Cluster Profiling
- Final Analysis and Outcome

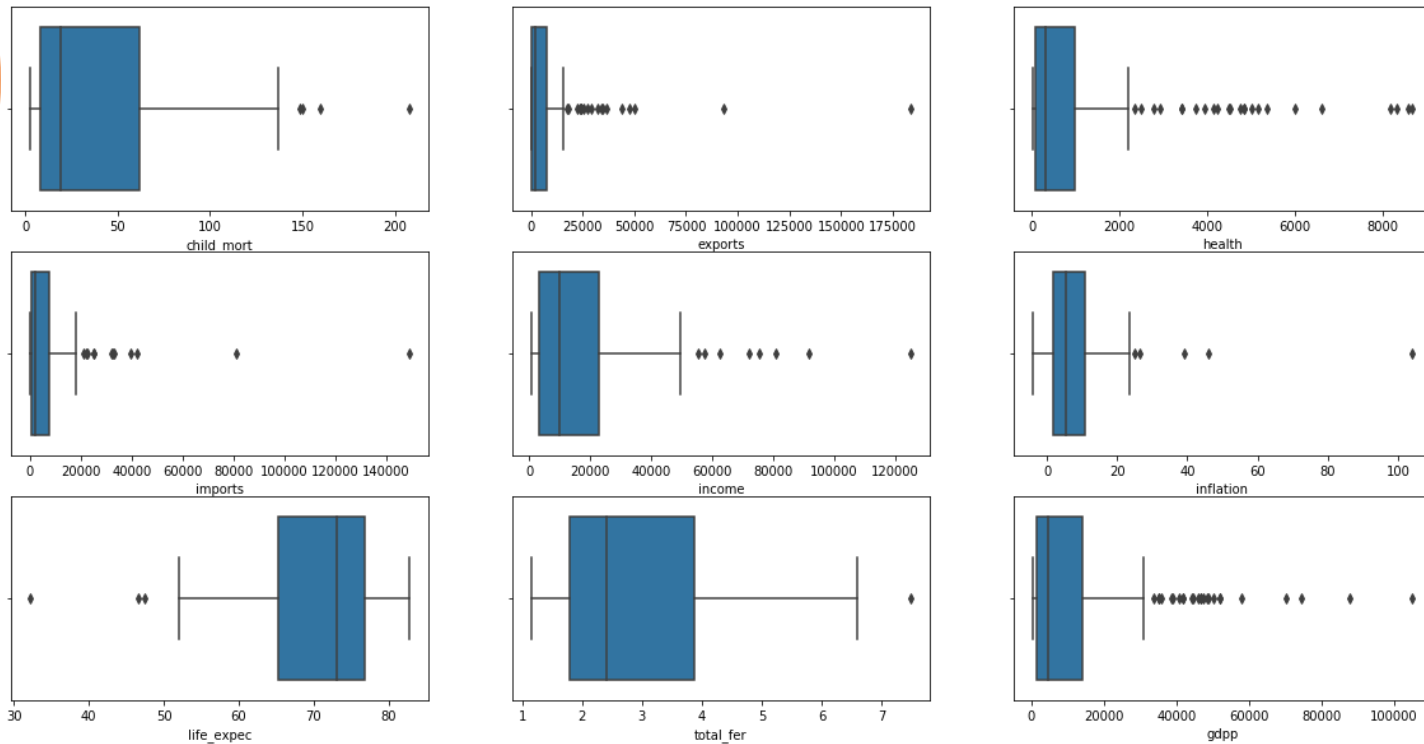


EDA Summary

- Univariate Analysis shown most of the African countries has
 - Lowest
 - Life Expectancy, GDP, Per Capita Income, Health, Imports, Exports
 - Highest
 - Child Mortality Rate, Inflation
- Heat Map analysis:-
 - child_mortality and life_expentency are highly correlated with correlation of -0.89
 - child_mortality and total_fertility are highly correlated with correlation of 0.85
 - imports and exports are highly correlated with correlation of 0.99
 - life_expentency and total_fertility are highly correlated with correlation of -0.76

Outlier Summary & Visualization

- We found outliers in almost all the features esp. in GDP and Health
- As we have only 167 countries, removing these outliers would shrink the shape of data and the under-developed countries which are in actual dire need may not contribute to the dataset
- The outliers were treated using soft capping of 1 – 99th percentile



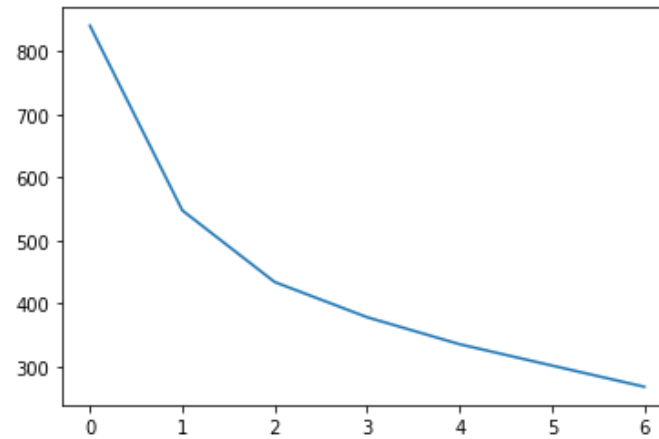


Hopkins Statistics

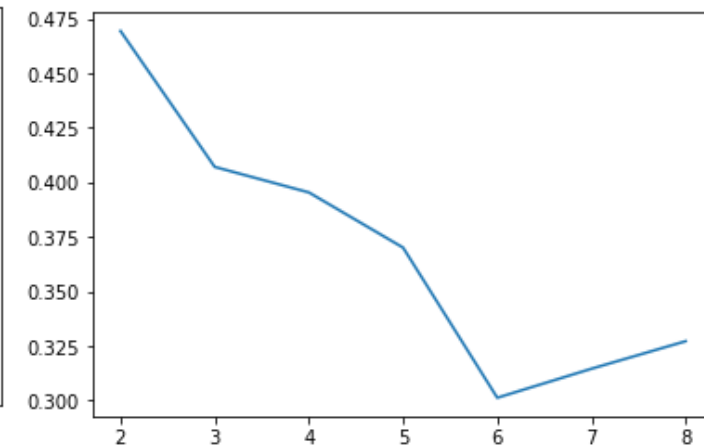
- A 'Hopkins Statistic' value close to 1 tends to indicate the data is highly clustered, random data will tend to result in values around 0.5, and uniformly distributed data will tend to result in values close to 0
- Hopkins Statistic of standardized scaled data came out to be 0.89
- Hopkins Statistic over 0.70 is a good score that indicated that the data is good for cluster analysis.

K-Means Clustering

- **K-means** is a centroid-based algorithm, or a distance-based algorithm, where we calculate the distances to assign a point to a cluster



Elbow/SSD curve

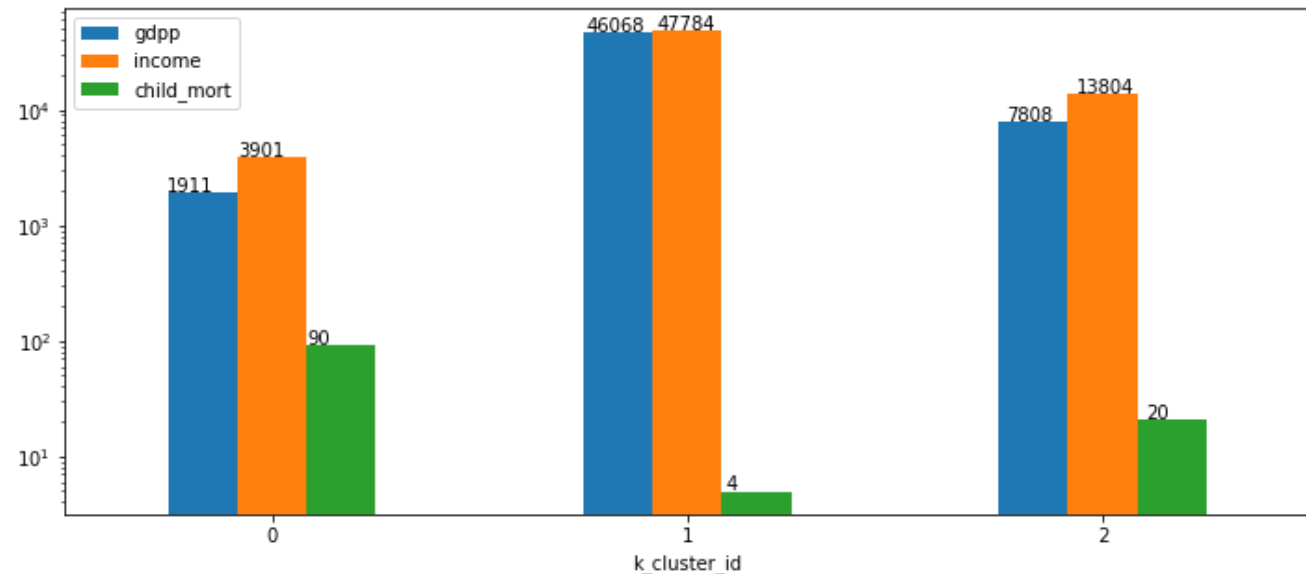


Silhouette Score

- As per elbow curve and silhouette score analysis 3 turned out to be the optimal cluster
- Also as per global economic factors dividing cluster into 3 seems logical and Final model was built using 3 clusters

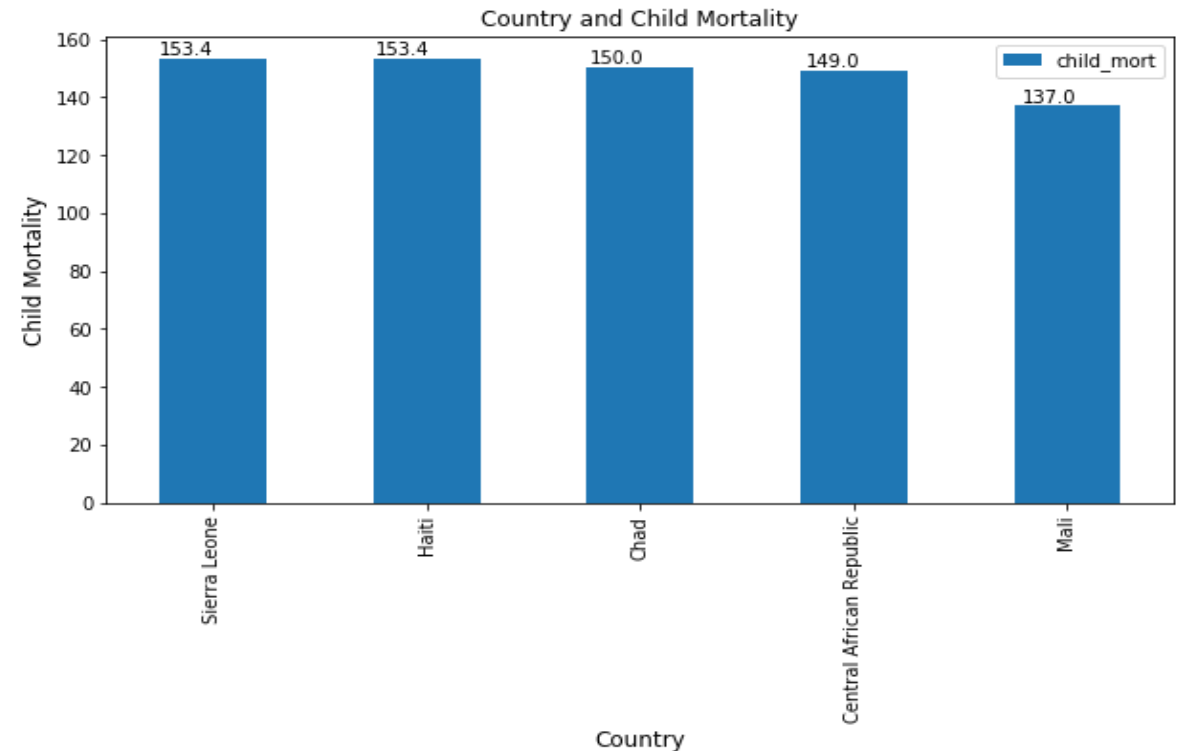
Cluster Summary (K-Means)

- Based on descriptive statistics 3 clusters are identified as-
 - Cluster 0 -> Under-Developed Countries
 - Cluster 1 -> Developed Countries
 - Cluster 2 -> Developing Countries
- Cluster 0 has low GDP, low income and highest child mortality rate
- All these factors made these cluster best candidate for financial aid from HELP NGO
- We also observed cluster 0 comprises of 29% of total data

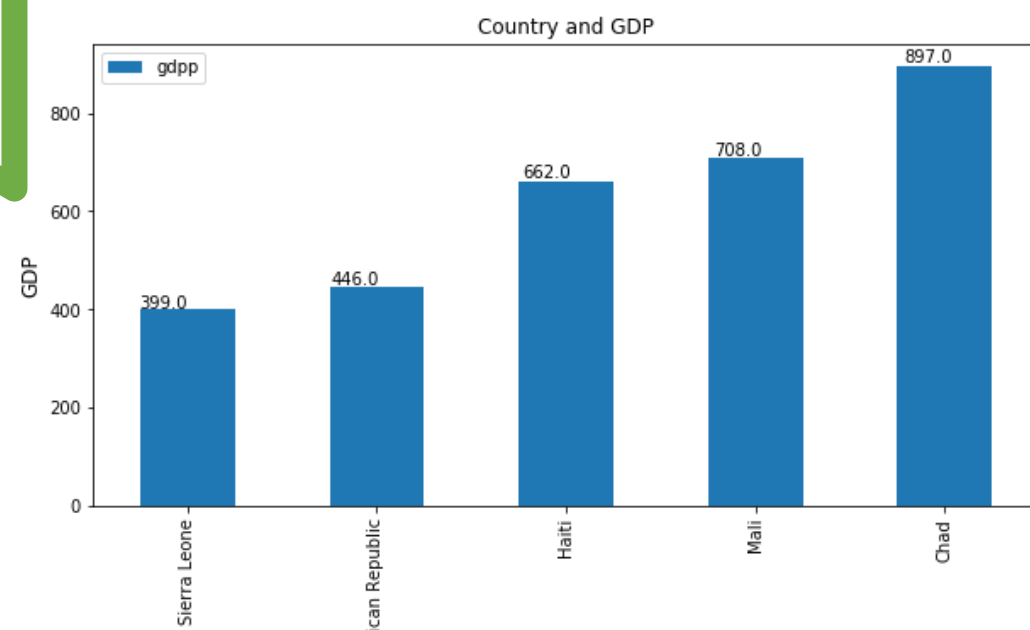


Cluster Profiling (K-Means)

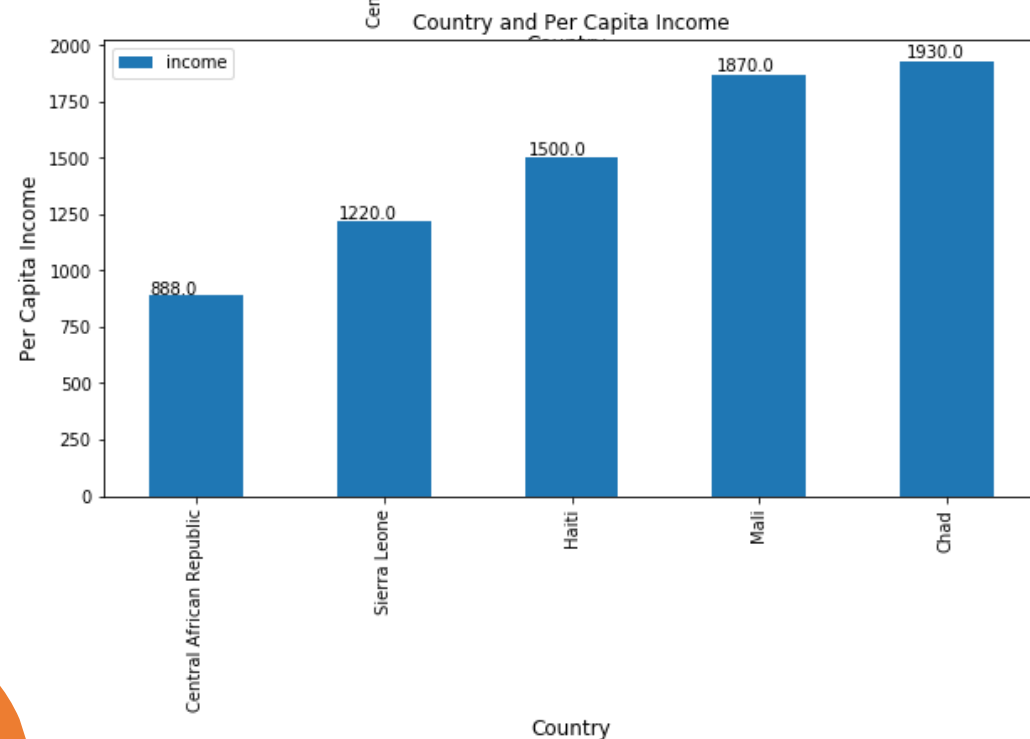
- Top 5 under-developed countries from cluster 0 in dire need of aid are
 - 0 Sierra Leone
 - 1 Haiti
 - 2 Chad
 - 3 Central African Republic
 - 4 Mali
- Visualization of Country Vs Child Mortality Rate, GDP, Income:-



Cluster Profiling (K-Means) – contd..



Country Vs GDP



Country Vs Income

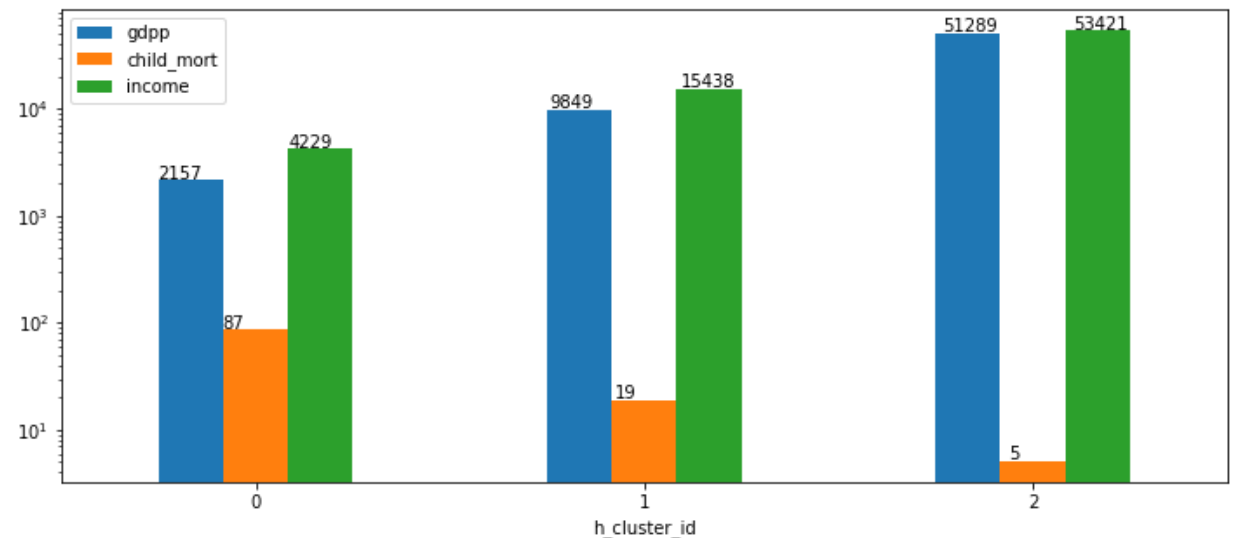


Hierarchical Clustering

- **Hierarchical cluster** analysis is an unsupervised **clustering** algorithm which seeks to build a hierarchy of clusters
 - Agglomerative – “Bottom-Up” Approach
 - Divisive – “Top-Down” Approach
- Linkage – It determines the distance between sets of observations
 - Single Linkage - the shortest distance between two points in each cluster
 - Complete Linkage - the longest distance between two points in each cluster

Cluster Summary (Hierarchical)

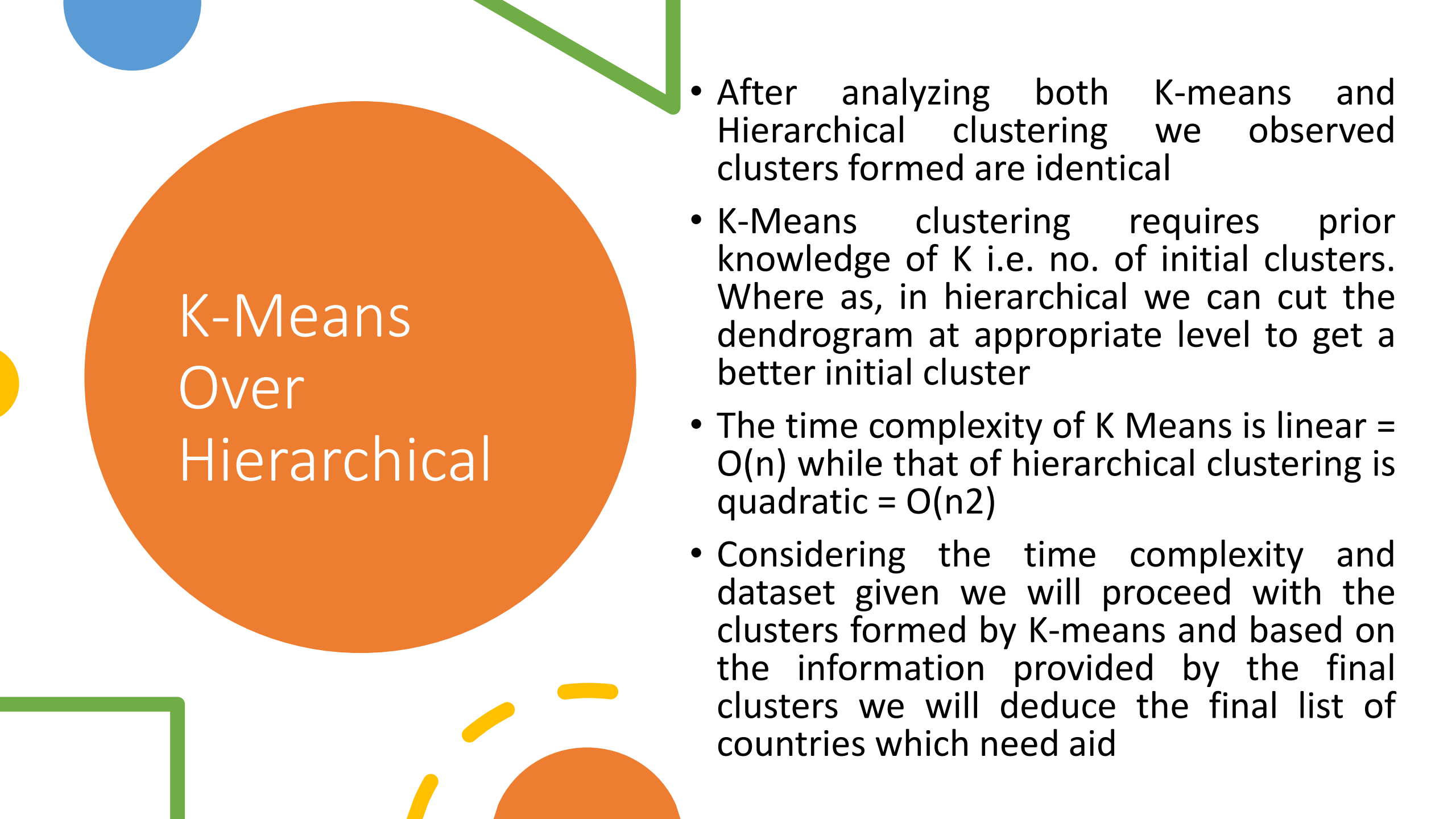
- Based on descriptive statistics 3 clusters are identified as-
 - Cluster 0 -> Under-Developed Countries
 - Cluster 1 -> Developing Countries
 - Cluster 2 -> Developed Countries
- Cluster 0 has low GDP, low income and highest child mortality rate
- All these factors made this cluster best candidate for financial aid from HELP NGO
- We also observed cluster 0 comprises of 50% of total data





Cluster Profiling (Hierarchical)

- We observed that Hierarchical clustering gave same set of countries as K-Means
- Top 5 under-developed countries from cluster 0 in dire need of aid are
 - 0 Sierra Leone
 - 1 Haiti
 - 2 Chad
 - 3 Central African Republic
 - 4 Mali



K-Means Over Hierarchical

- After analyzing both K-means and Hierarchical clustering we observed clusters formed are identical
- K-Means clustering requires prior knowledge of K i.e. no. of initial clusters. Where as, in hierarchical we can cut the dendrogram at appropriate level to get a better initial cluster
- The time complexity of K Means is linear = $O(n)$ while that of hierarchical clustering is quadratic = $O(n^2)$
- Considering the time complexity and dataset given we will proceed with the clusters formed by K-means and based on the information provided by the final clusters we will deduce the final list of countries which need aid



Final Analysis

- We concluded top 5 countries from final cluster (under-developed countries) based on GDP, Child Mortality Rate and Per Capita Income
 1. Sierra Leone
 2. Haiti
 3. Chad
 4. Central African Republic
 5. Mali
- Below is the ordered we filtered the list of country names:
 - Lowest GDP
 - Lowest Income
 - Highest Child Mortality Rate

Final Statistics of Recommended Countries

	gdpp	income	child_mort
count	5.000000	5.000000	5.000000
mean	622.400000	1481.600000	148.560000
std	203.301008	439.616651	6.75929
min	399.000000	888.000000	137.000000
25%	446.000000	1220.000000	149.000000
50%	662.000000	1500.000000	150.000000
75%	708.000000	1870.000000	153.400000
max	897.000000	1930.000000	153.400000

