Clustering Assignment

Submitted by:

Abhinav Choudhary (PGDDS C-24)

Problem Statement

- 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.

Business Objective

- The business wants to categorize the countries using some socio-economic and health factors that determine the overall development of the country. We need to suggest CEO the countries which are in the direst need of financial aid.

Analysis Approach

Understanding the Data - EDA & Data Visualization

- Import & understand the data at a high level
- Transform few columns to their absolute data
- EDA – Univariate & Bivariate Analysis
- Outlier Analysis & Treatment

Hopkin's Test

- To check the Clustering Tendency of the dataset

Scaling

- To Standardize all the continuous variables.

K-Means Clustering

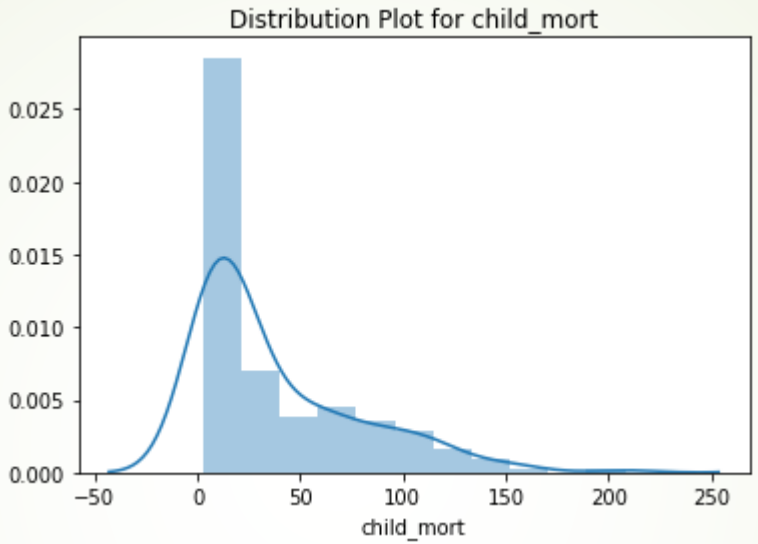
- Determine the best value of K : SSD & Silhoutte Analysis
- K-Means Analysis
- Cluster Profiling
- Identifying the top Countries (with direst need of financial aid)

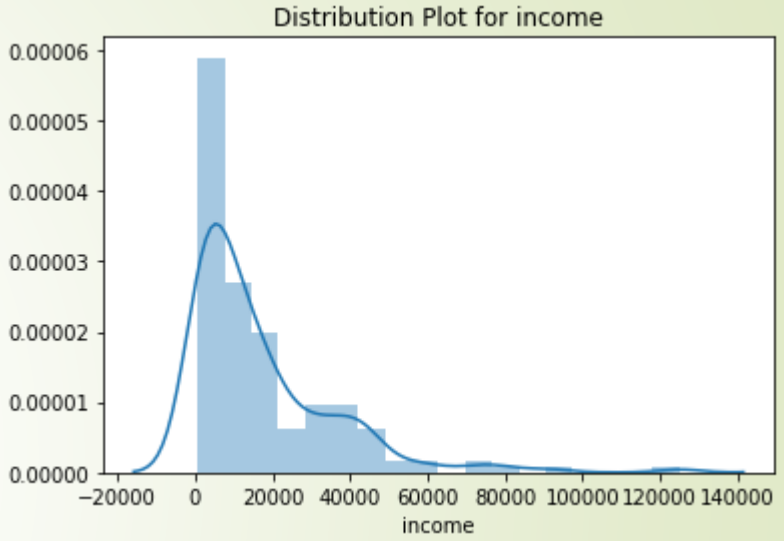
Hierarchical Clustering

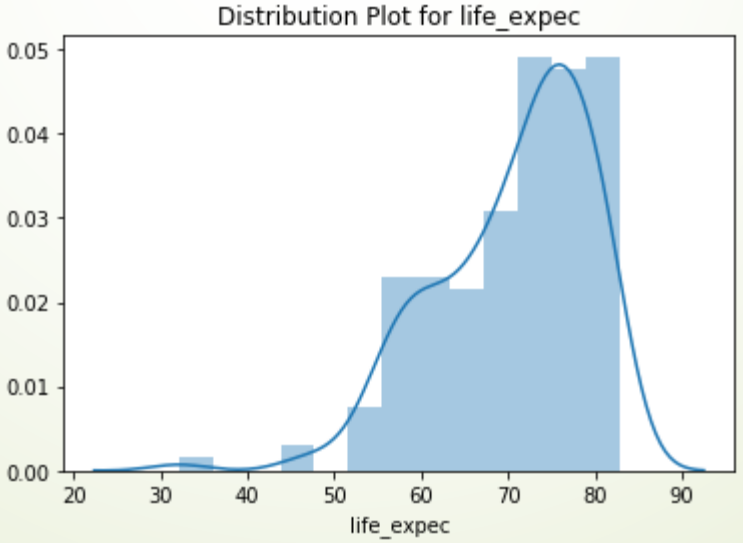
- Clustering using Single & Complete linkage
- Cluster Profiling
- Identifying the top Countries (with direst need of financial aid)

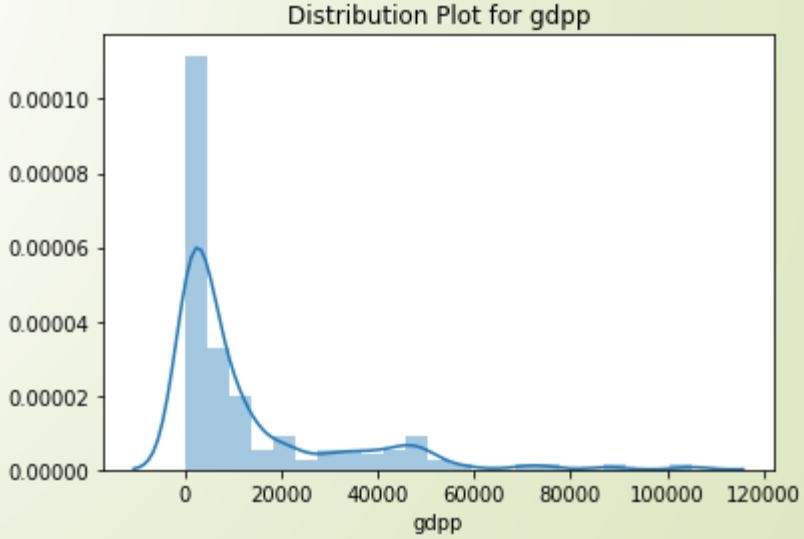
EDA – Univariate & Bivariate Analysis

- Below are the inferences obtained from Univariate Analysis:
  - Except for "life_expec"(which is left-skewed distributed), rest all the features are right-skewed distributed.
  - The distribution of "income", "gdpp" & "child_mort" is very much similar which indicates a sign of having internal grouping. Therefore, these three features would be more useful for us to do the Cluster Profiling.



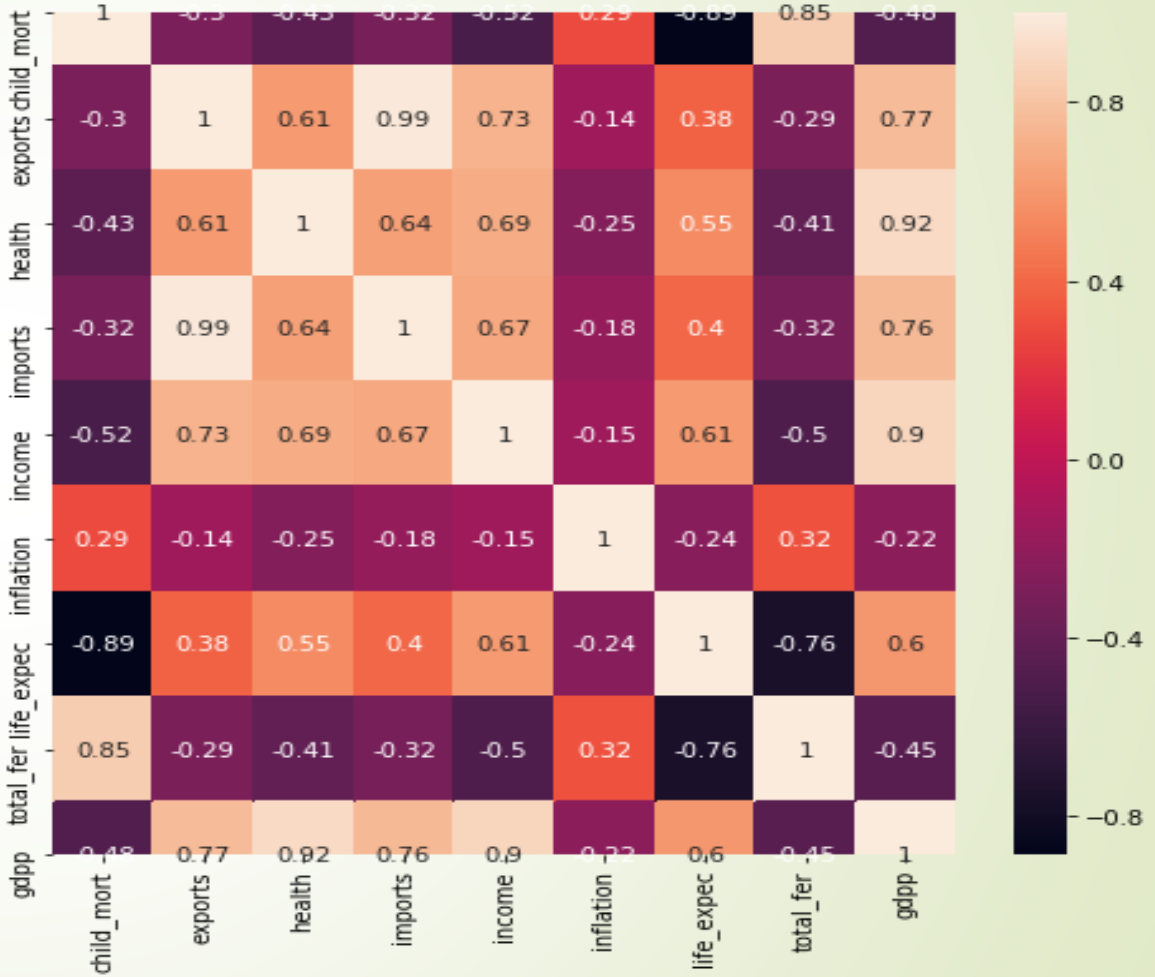






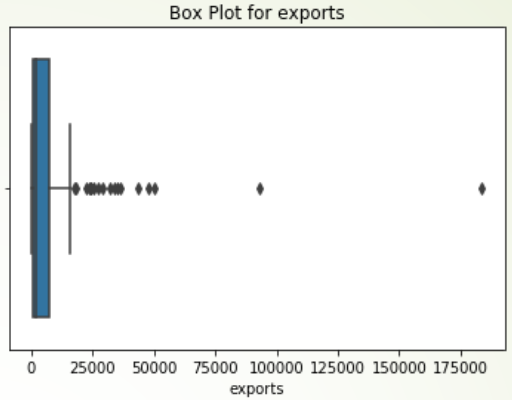
EDA – Univariate & Bivariate Analysis

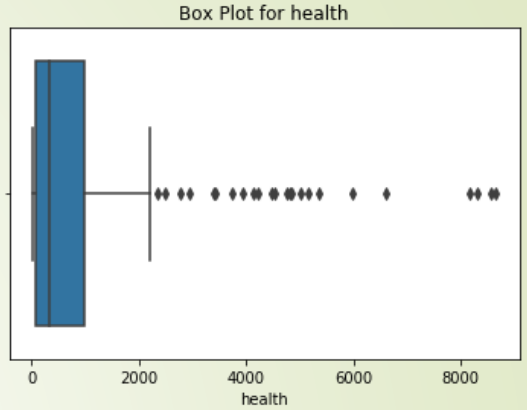
- From the Correlation Matrix, it can be observed that the below feature pairs have a high positive linear correlation:
  - gdpp & income
  - gdpp & health
  - child_mort & total_fer
  - exports & imports
- "child_mort" & "life_expec" have a high negative correlation which is quite expected here.

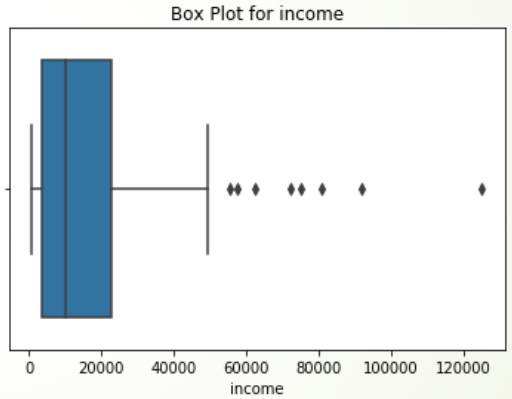


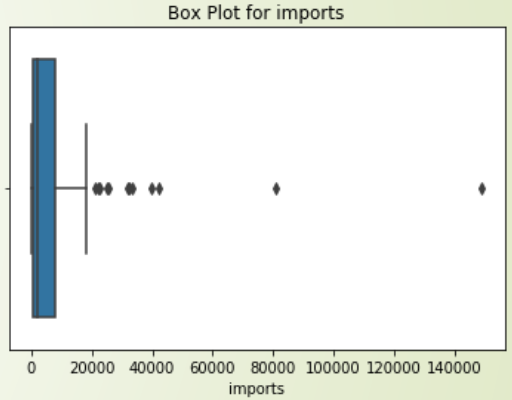
Outlier Analysis & Treatment

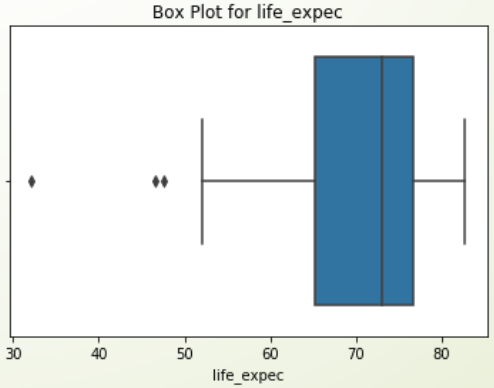
- From the Box plot it can be observed that except for "life_expec" (have outliers present at the lower end), all the other features have outliers present at the higher end.
- As we know the business objective here is to identify the countries which are in the direst need of financial aid, therefore, we can perform capping of the outlier countries (from each of the features) which represents financially stable countries. This way we will not be disturbing any data points which are the representative of poor countries.
- Therefore, here we have performed the soft capping of the upper range outliers of the below features:
  - exports; health; imports; income & gdpp

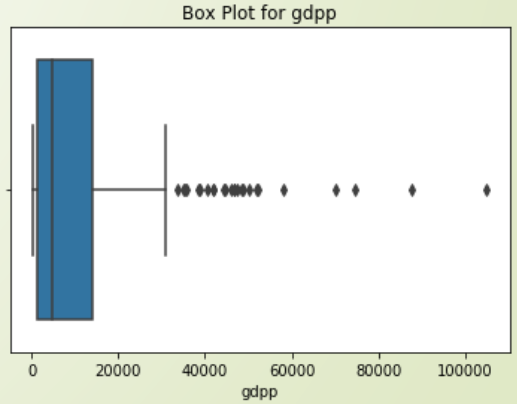










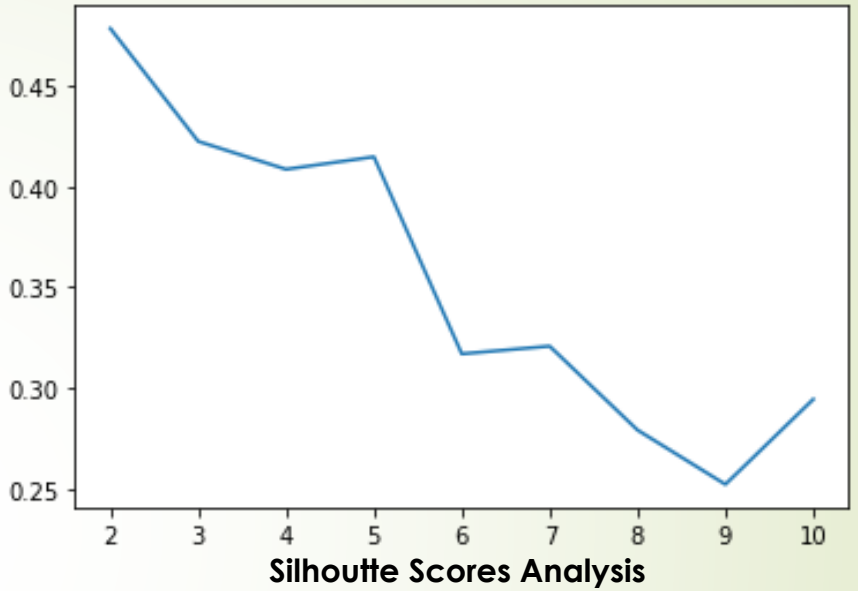


Hopkin's Test

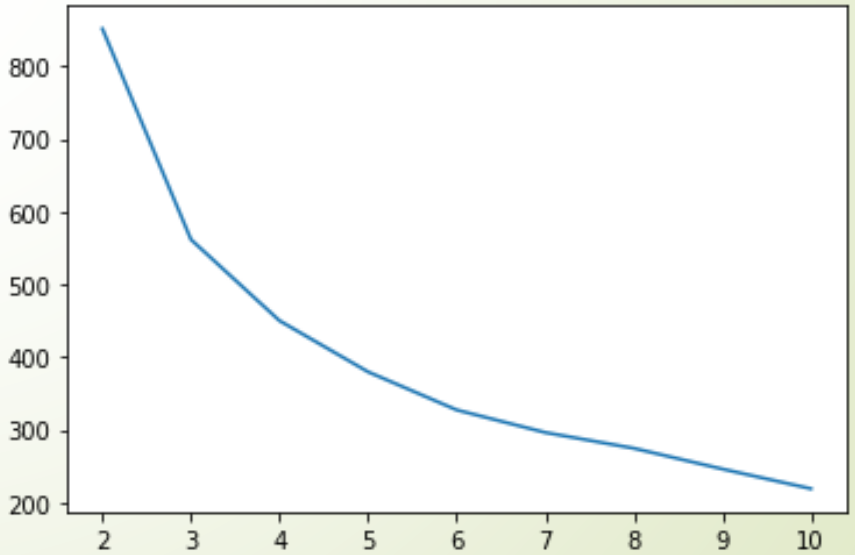
- Through Hopkins Test, we tried to understand how different our data is from randomly scattered data.
- We performed this test multiple times to analyze the results & take necessary actions.
- So, from the Hopkin's Test, we found that the average score has come out to be around 90% which is an excellent indicator of higher clustering tendency of our dataset.
- In other words, this test also represents that our dataset is 90% different from the randomly scattered data.
  The higher the value of dissimilarity, the more apt is our data for clustering.

K-Means Clustering: Silhoutte & SSD Analysis

- **Silhoutte Analysis**: From the Silhouette score analysis, we can see that the maximum score (after k=2) is obtained at **k=3**
- **SSD Analysis**: From the SSD analysis, we can see that the slope of the line (after k=2) has a maximum delta/change at **k=3**.
- Therefore, from the two analysis(SSD & silhouette), we can conclude that the **best value of k must be 3**



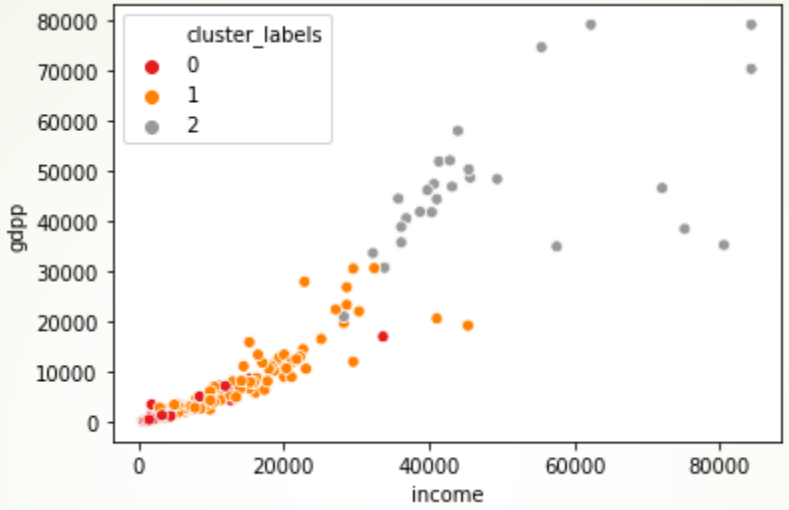
Silhoutte Scores Analysis

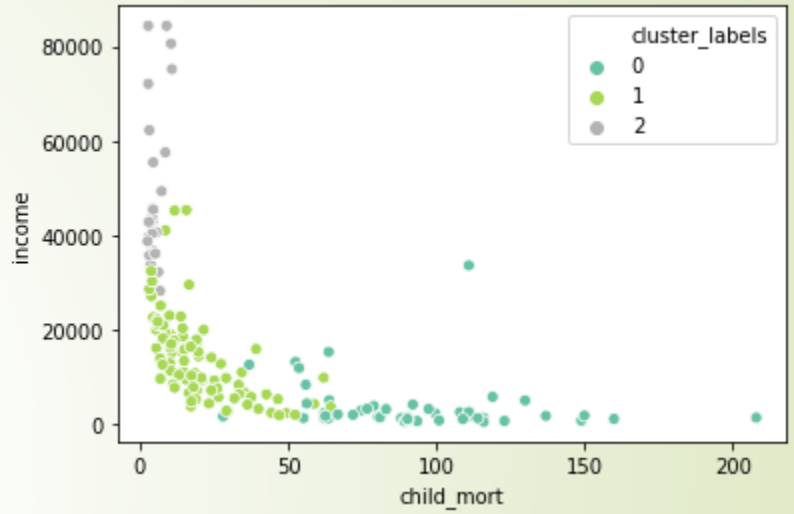


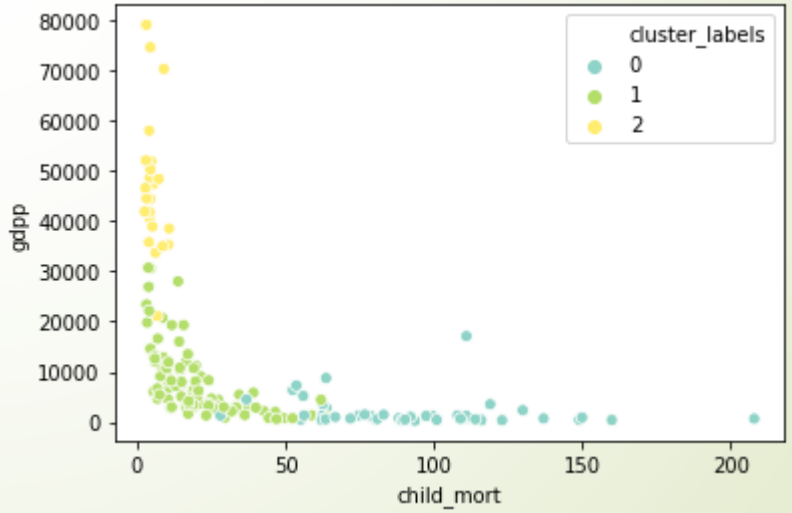
SSD Analysis

K-Means Clustering: Cluster Profiling

- Here, we visualized the clusters using the scatter plot with respect to the below features (in pairs):
  - gdpp
  - child_mort
  - income
- From the scatter plots here, the 3 clusters are neat & visible to us.
  We can also clearly identify the groups of countries based on the level of their financial stability.
  With that in mind, let's proceed & analyze further in Cluster Profiling

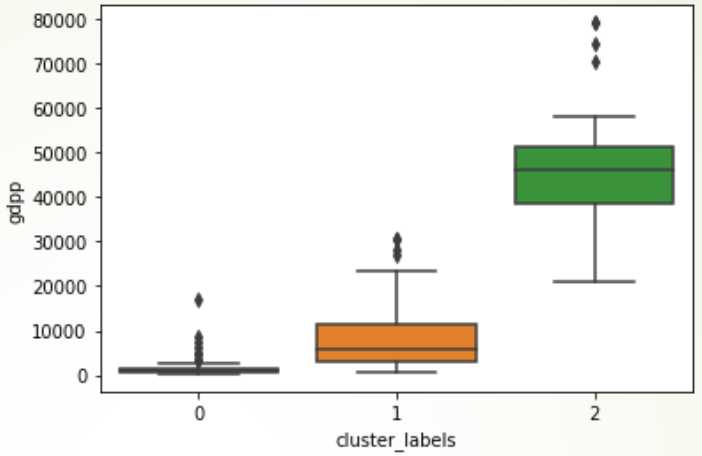


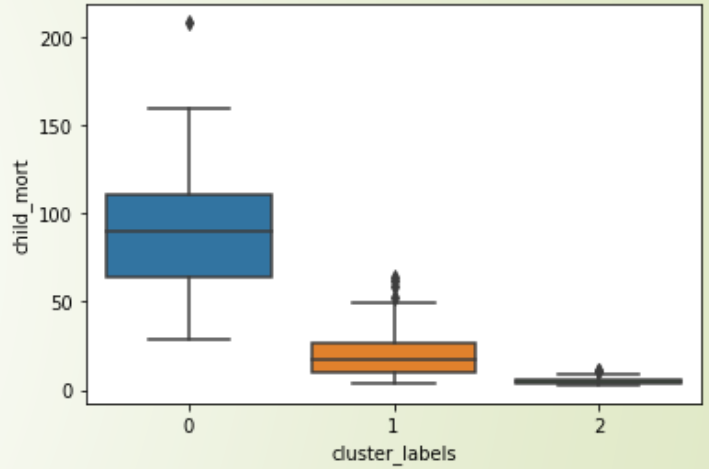


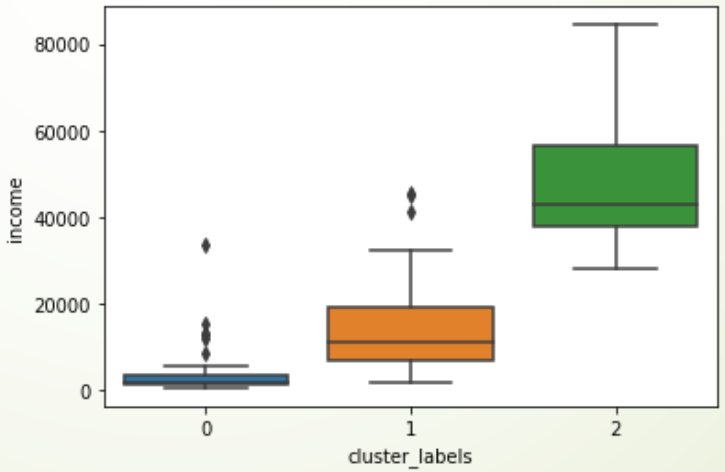


K-Means Clustering: Cluster Profiling (cont'd)

- Next, here we performed the cluster profiling through Box plots with respect to the below features:
  - gdpp
  - child_mort
  - income
- Our objective here is to identify that cluster of countries which are in direst need of financial aid. That is the countries with LOW GDPP, LOW Income & HIGH Child Mortality.
- From the box plots, it is clearly evident that our target cluster is the one with cluster_labels=0. This cluster represents the poor countries with LOW GDPP, LOW Income & HIGH Child Mortality.







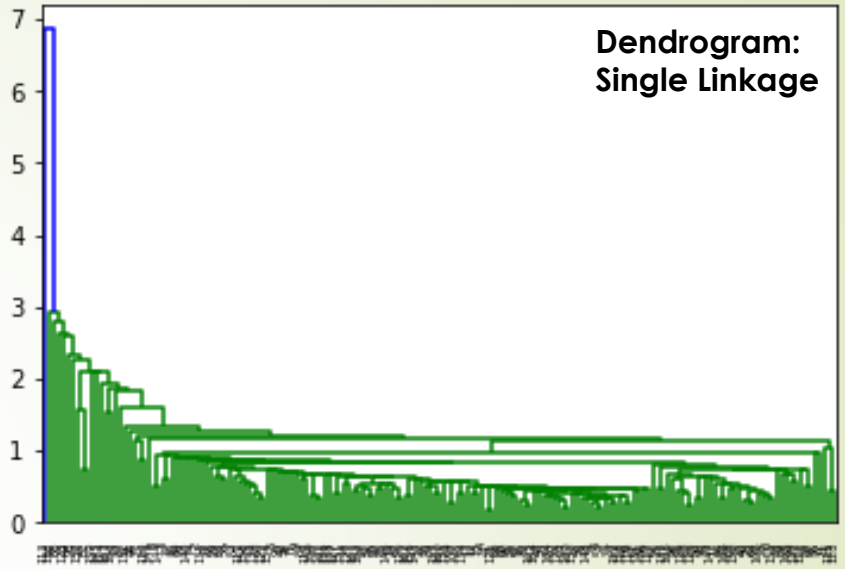
K-Means Clustering: Identifying the top Countries (with direst need of financial aid)

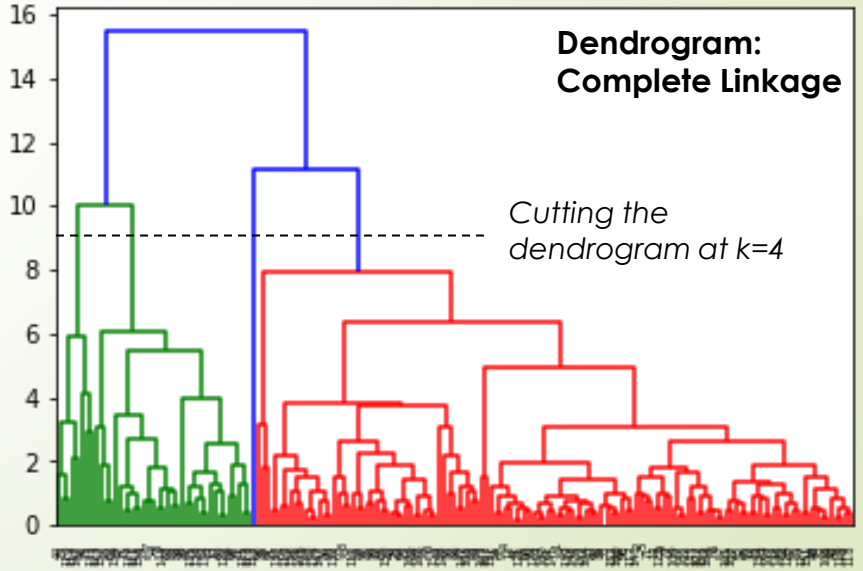
- Post cluster Profiling, we identified & further analysed the target cluster (of poor countries).
- After sorting the cluster in the given order we found the top countries with direst need of financial aid
- Below is the list such countries

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	cluster_labels
26	Burundi	93.6	20.6052	26.7960	90.552	764.0	12.30	57.7	6.26	231.0	0
88	Liberia	89.3	62.4570	38.5860	302.802	700.0	5.47	60.8	5.02	327.0	0
37	Congo, Dem. Rep.	116.0	137.2740	26.4194	165.664	609.0	20.80	57.5	6.54	334.0	0
112	Niger	123.0	77.2560	17.9568	170.868	814.0	2.55	58.8	7.49	348.0	0
132	Sierra Leone	160.0	67.0320	52.2690	137.655	1220.0	17.20	55.0	5.20	399.0	0
93	Madagascar	62.2	103.2500	15.5701	177.590	1390.0	8.79	60.8	4.60	413.0	0
106	Mozambique	101.0	131.9850	21.8299	193.578	918.0	7.64	54.5	5.56	419.0	0
31	Central African Republic	149.0	52.6280	17.7508	118.190	888.0	2.01	47.5	5.21	446.0	0
94	Malawi	90.5	104.6520	30.2481	160.191	1030.0	12.10	53.1	5.31	459.0	0
50	Eritrea	55.2	23.0878	12.8212	112.306	1420.0	11.60	61.7	4.61	482.0	0

Hierarchical Clustering

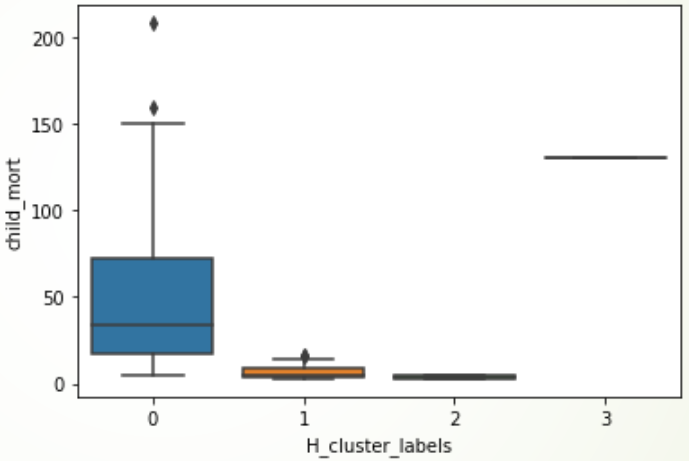
- Here we started with performing the Hierarchical Clustering through the below two linkage methods:
  - **Single Linkage**: The dendrogram obtained is not clear to make any cut for clustering. Hence, we will go with dendrogram with complete linkage.
  - **Complete Linkage**: Here, from the above dendrogram(with complete linkage) we can either cut it at k=3 or at k=4.
- From the Complete Linkage dendrogram, we decided to go with **k=4** so that we can compare our results with that of K-means clustering outcomes.
- Here are the Dendrograms obtained from each of the linkage methods ->

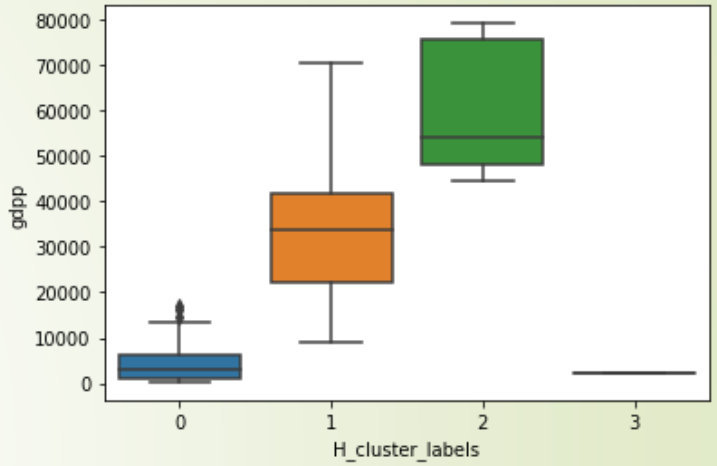


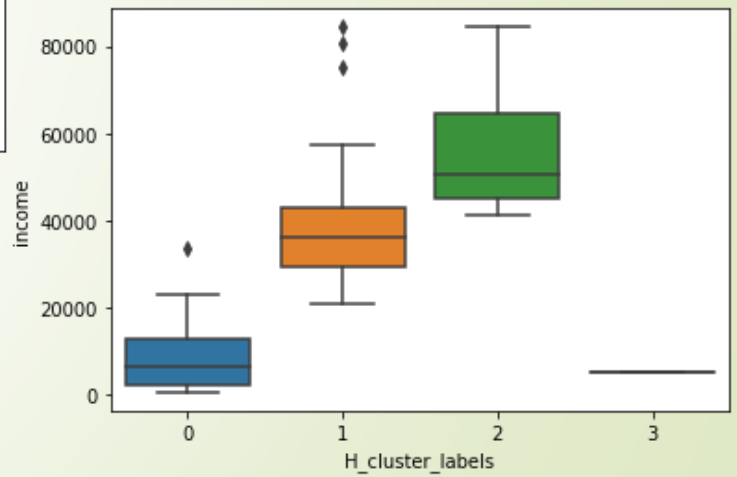


Hierarchical Clustering: Cluster Profiling

- From this cluster profiling exercise through the box plots, it is clearly evident that our target cluster is the one with H_cluster_labels=0. This cluster represents the poor countries with LOW GDPP, LOW Income & HIGH Child Mortality.
- Here, we can ignore the last cluster with H_cluster_labels=3 as it only has one record
- With this, we can even conclude that our dataset is best suited for having 3 clusters ()







Hierarchical Clustering: Identifying the top Countries (with direst need of financial aid)

- Post cluster Profiling, we identified & further analysed our target cluster (of poor countries).
- After sorting the cluster in the given order we found the top countries with direst need of financial aid
- Below is the list such countries

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	cluster_labels	H_cluster_labels
26	Burundi	93.6	20.6052	26.7960	90.552	764.0	12.30	57.7	6.26	231.0	0	0
88	Liberia	89.3	62.4570	38.5860	302.802	700.0	5.47	60.8	5.02	327.0	0	0
37	Congo, Dem. Rep.	116.0	137.2740	26.4194	165.664	609.0	20.80	57.5	6.54	334.0	0	0
112	Niger	123.0	77.2560	17.9568	170.868	814.0	2.55	58.8	7.49	348.0	0	0
132	Sierra Leone	160.0	67.0320	52.2690	137.655	1220.0	17.20	55.0	5.20	399.0	0	0
93	Madagascar	62.2	103.2500	15.5701	177.590	1390.0	8.79	60.8	4.60	413.0	0	0
106	Mozambique	101.0	131.9850	21.8299	193.578	918.0	7.64	54.5	5.56	419.0	0	0
31	Central African Republic	149.0	52.6280	17.7508	118.190	888.0	2.01	47.5	5.21	446.0	0	0
94	Malawi	90.5	104.6520	30.2481	160.191	1030.0	12.10	53.1	5.31	459.0	0	0
50	Eritrea	55.2	23.0878	12.8212	112.306	1420.0	11.60	61.7	4.61	482.0	0	0

Summary & Key Takeaway

- So finally, we have observed that from both **K-Means** & **Hierarchical** Clustering, although the process & clusters formed were different, but the end-result obtained(the list of countries with the direst need of financial aid) is the same.
- The following are the countries which are in the direst need of financial aid by considering socioeconomic factors into consideration:

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	57.7	6.26	231.0
Liberia	89.3	62.4570	38.5860	302.802	700.0	5.47	60.8	5.02	327.0
Congo, Dem. Rep.	116.0	137.2740	26.4194	165.664	609.0	20.80	57.5	6.54	334.0
Niger	123.0	77.2560	17.9568	170.868	814.0	2.55	58.8	7.49	348.0
Sierra Leone	160.0	67.0320	52.2690	137.655	1220.0	17.20	55.0	5.20	399.0
Madagascar	62.2	103.2500	15.5701	177.590	1390.0	8.79	60.8	4.60	413.0
Mozambique	101.0	131.9850	21.8299	193.578	918.0	7.64	54.5	5.56	419.0
Central African Republic	149.0	52.6280	17.7508	118.190	888.0	2.01	47.5	5.21	446.0
Malawi	90.5	104.6520	30.2481	160.191	1030.0	12.10	53.1	5.31	459.0
Eritrea	55.2	23.0878	12.8212	112.306	1420.0	11.60	61.7	4.61	482.0

Thank You! ☺