Clustering Assignment

- Denis Roystan Dalmeida

Problem Statement

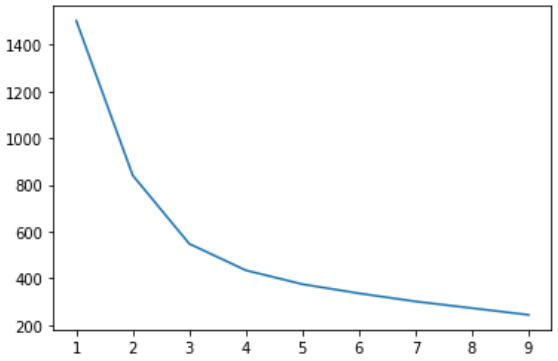
- HELP International is an international humanitarian NGO that is committed to fighting poverty and providing the people of backward countries with basic amenities and relief during the time of disasters and natural calamities.
- HELP have been able to raise $ 10 million in the recent funding programmes.
- CEO of the NGO needs to decide how to use raised money strategically and effectively so that the countries that are in the direst need of aid would get the first priority from organization.
- As a Data Analyst, needs to categorise the countries using some socio-economic and health factors that determine the overall development of the country. Then as an analyst need to suggest the countries which the CEO needs to focus on the most. Make sure to report back at least 5 countries which are in direst need of aid from the analysis.

Analysis Approach

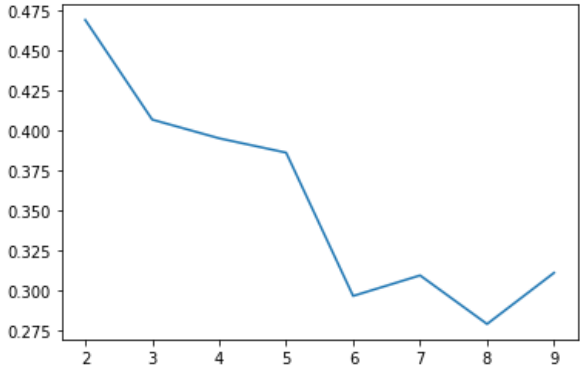
- Starting off with necessary data inspection and EDA tasks suitable for dataset - data cleaning, univariate analysis, bivariate analysis etc.
- Performing the Outlier Analysis on the dataset. If Outliers are present decide whether to treat and keep them or remove them depending on the results you get. One can do capping of outliers to the 99 percentile.
- Performing both the types of clustering on data – K-means and Hierarchical Clustering to create clusters
- Analyzing the clusters and identify the ones which are in dire need of aid. We have analyzed the clusters by comparing these three variables - [**gdpp**, **child_mort** and **income**] vary for each cluster of countries to recognize and differentiate the clusters of developed countries from the clusters of under-developed countries.
- Performing visualizations on the clusters that have been formed.
- Choosing above one method so that to report the final list of countries that are in dire need of aid and also highlighting the first 5 countries from that list based on socio-economic and health factor [**gdpp**, **child_mort** and **income**]

Results of Clustering Model

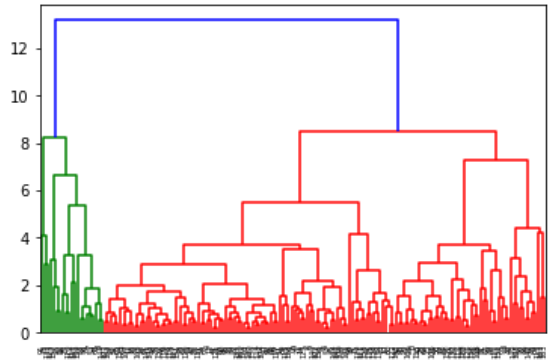
- The clustering model outputs the 3 cluster ids (0,1,2) for K-means and 2 cluster ids/labels (0,1) for Hierarchical Clustering based on no of cluster selection.



Elbow curve method



Silhouette score

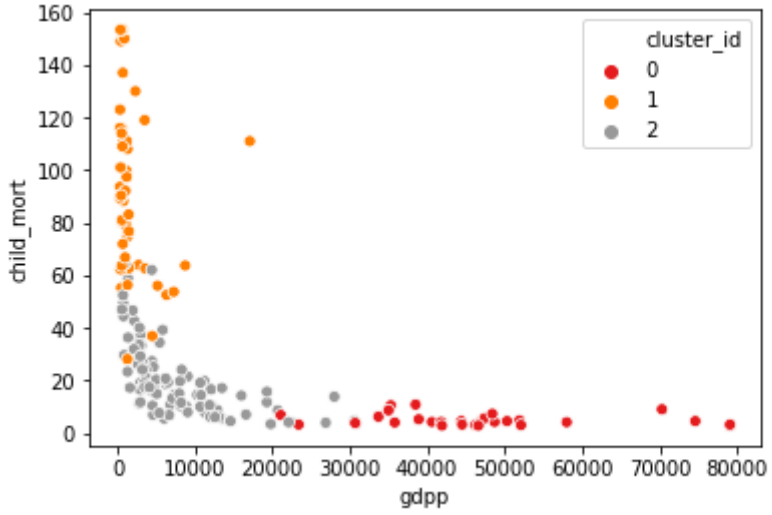


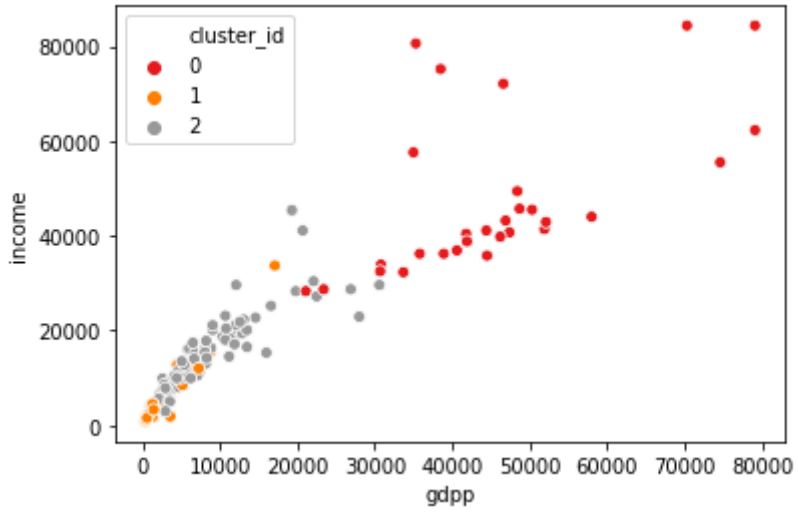
Complete Linkage (Dendrogram)

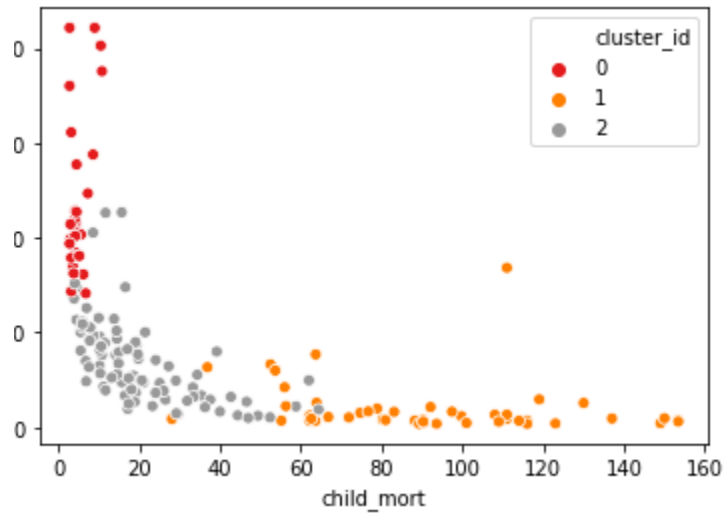
- Based on the graphs above one can say that k=3 for K-means Clustering and K=2 for Hierarchical Clustering.

Visualisations

- K-MEANS Clustering using Scatter Plot based on three factors selected randomly at two to plot against X and Y [gdpp, child_mort, income]

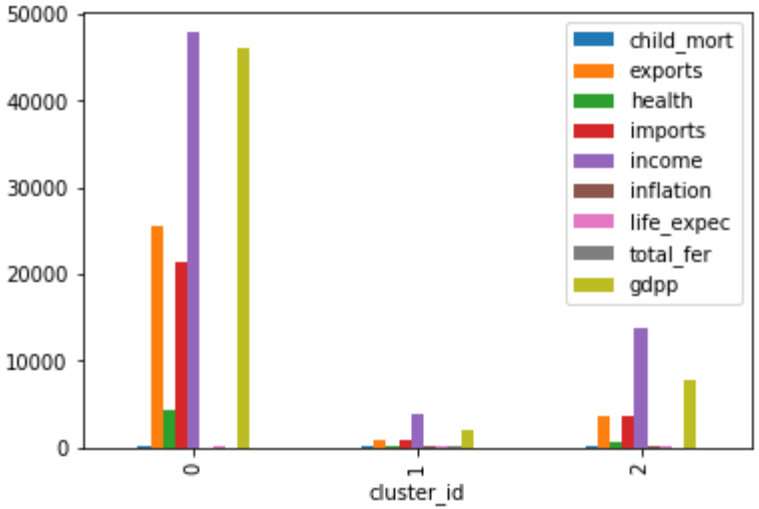




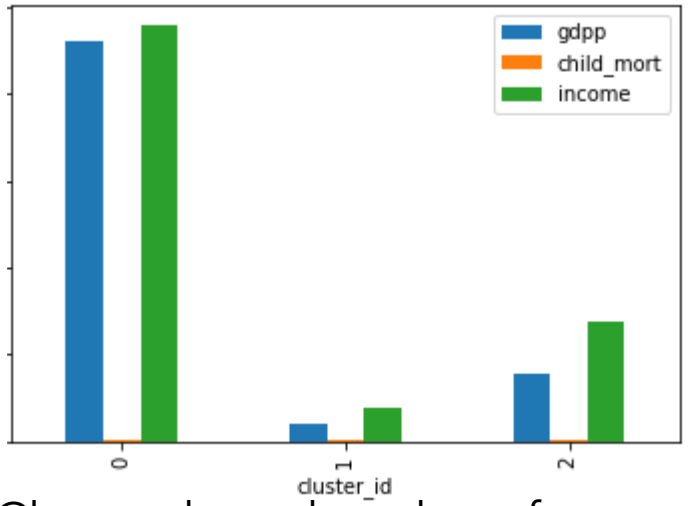


Visualisations

- K-MEANS Clustering: Analysing clusters based on plotted bar graphs with X as cluster-id and Y as factors listed below.



Clusters based on all factors

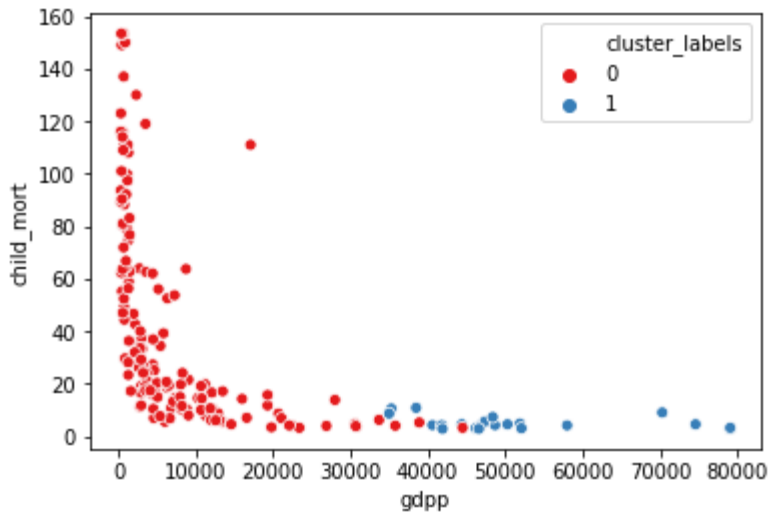


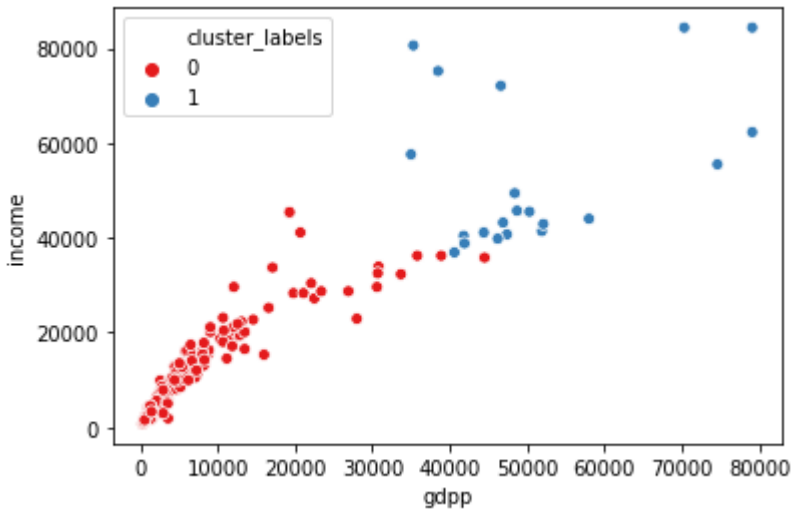
Clusters based on three factors [gdpp, child_mort, income]

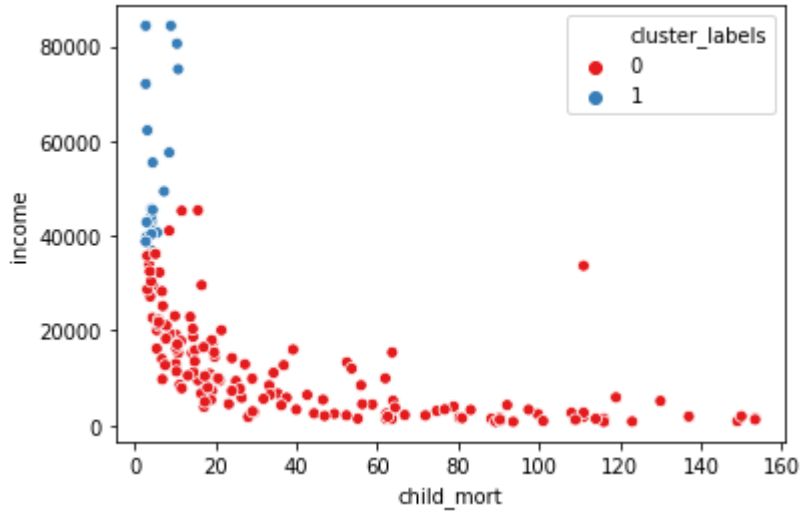
Looks like countries with cluster-id "1" are in dire need of aid based on K-means clustering

Visualisations

- Hierarchical Clustering: Using Scatter Plot based on three factors selected randomly at two to plot against X and Y [gdpp, child_mort, income]

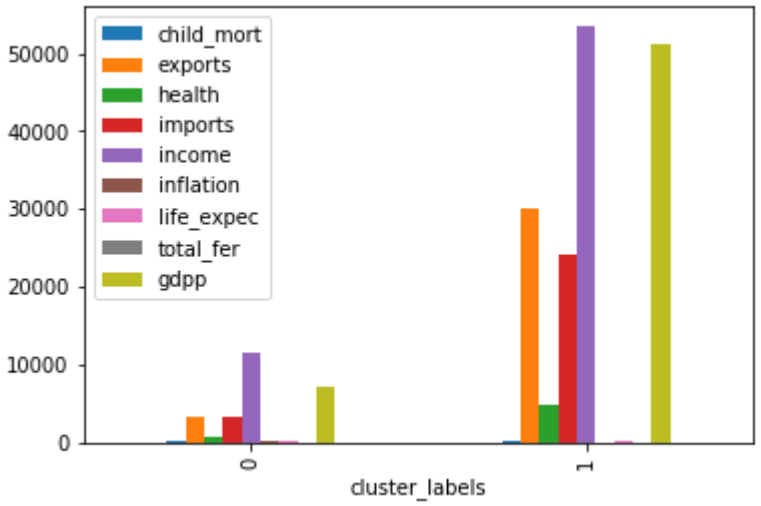


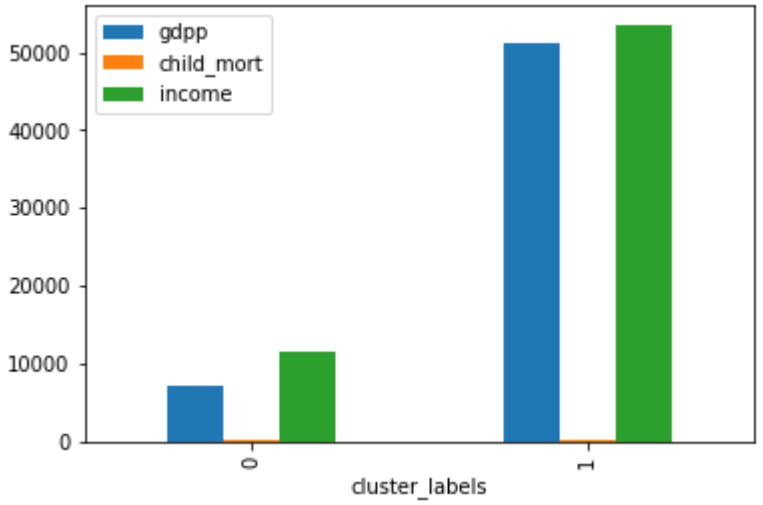




Visualisations

- Hierarchical Clustering: Analysing clusters based on plotted bar graphs with X as cluster-labels and Y as factors listed below.





Looks like countries with cluster-id "0" are in dire need of aid based on Hierarchical clustering

Final list of countries

Final list using
K-Means Clustering

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	cluster_id
88	Liberia	89.3	62.457000	38.5860	302.80200	742.24	5.47	60.8	5.0200	331.62	1
26	Burundi	93.6	22.243716	26.7960	104.90964	764.00	12.30	57.7	6.2600	331.62	1
37	Congo, Dem. Rep.	116.0	137.274000	26.4194	165.66400	742.24	20.80	57.5	6.5400	334.00	1
112	Niger	123.0	77.256000	17.9568	170.86800	814.00	2.55	58.8	6.5636	348.00	1
132	Sierra Leone	153.4	67.032000	52.2690	137.65500	1220.00	17.20	55.0	5.2000	399.00	1

Final list using
Hierarchical Clustering

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	cluster_labels
88	Liberia	89.3	62.457000	38.5860	302.80200	742.24	5.47	60.8	5.0200	331.62	0
26	Burundi	93.6	22.243716	26.7960	104.90964	764.00	12.30	57.7	6.2600	331.62	0
37	Congo, Dem. Rep.	116.0	137.274000	26.4194	165.66400	742.24	20.80	57.5	6.5400	334.00	0
112	Niger	123.0	77.256000	17.9568	170.86800	814.00	2.55	58.8	6.5636	348.00	0
132	Sierra Leone	153.4	67.032000	52.2690	137.65500	1220.00	17.20	55.0	5.2000	399.00	0

Final list of countries

- Based on the previous slide results using clustering methods and the thorough analysis of the countries data we have reached to the final list of countries along with Top 5 countries which are in dire need of aid.
- Top 5 countries which are in direst need of aid
  - Liberia
  - Burundi
  - Congo, Dem. Rep.
  - Niger
  - Sierra Leone