Clustering Assignment
(K-Means & Hierarchical Clustering)

By: Vishal D Mehta

Background

Objective:

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 programmes, 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:

Our job is to categorise the countries using some socio-economic and health factors that determine the overall development of the country. We need to suggest the countries which the CEO needs to focus on the most.

Analysis-methodology

Data collection and cleaning

- Importing the data
- Identifying the data quality issues and cleaning the data

Data Transformation

- Transforming the variables in appropriate form for analysis .Outliers analysis is avoided as it would hamper in achieving the end business goal.

Visualizing the data

- Visualizing few original data variables to look for any pattern or correlation.

Scaling the data

- Standardizing all the continuous variables.

Hopkins Statistics

- To check if data has tendency to form clusters

Hierarchical Clustering

- Identify the 'n' via dendrogram.
- Forming n – clusters on the dataset
- Visualizing the clusters with various variables
- Analysing the clusters
- Identifying the countries which requires aid.

Analysis-methodology Cont...

K means clustering

- Identify the 'k' by silhouette analysis and sum of squared distances graph.
- Forming n – clusters on the dataset
- Visualizing the clusters with various variables
- Analyzing the clusters
- Identifying the countries which requires aid.

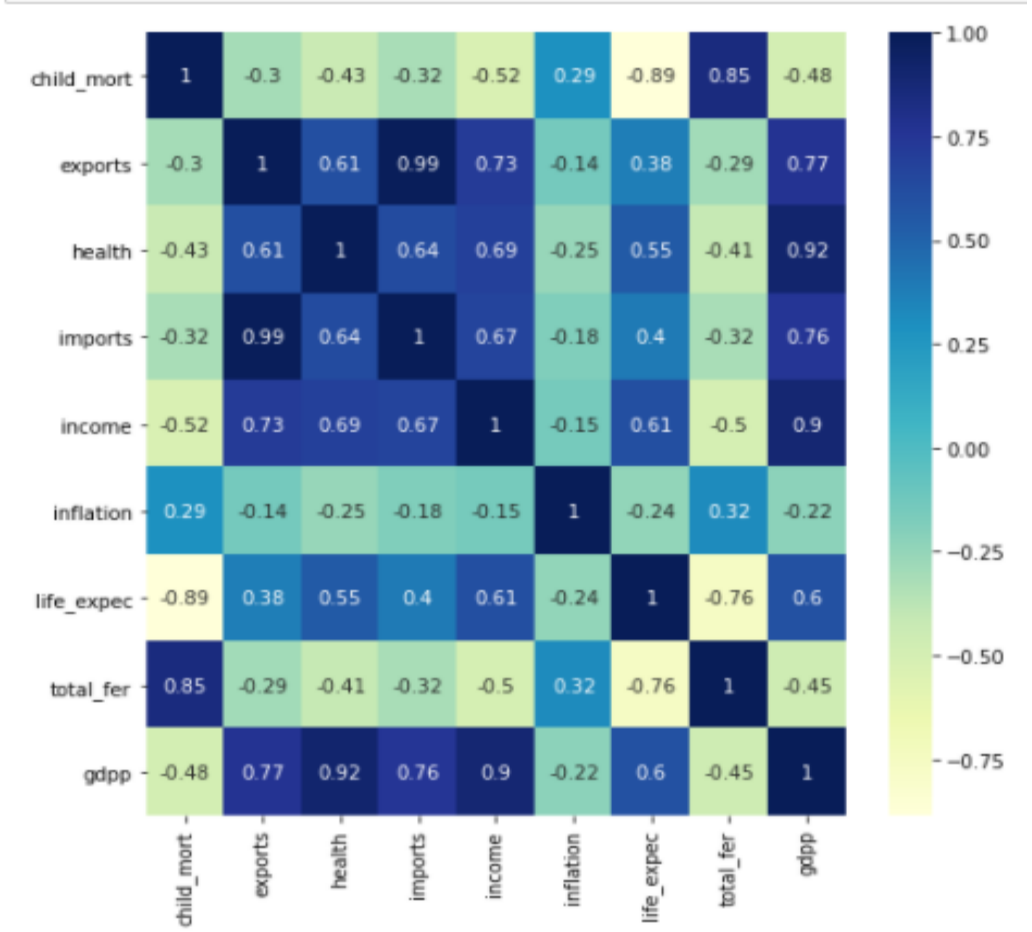
Making Final Cluster of Countries needing aid

- A final cluster was made for analysis which had set of countries which needed the financial aid

Decision Making

- Final cluster of Under Developed Countries where analysed and out of that top 10 countries where selected for which the NGO can aim for Financial aid .

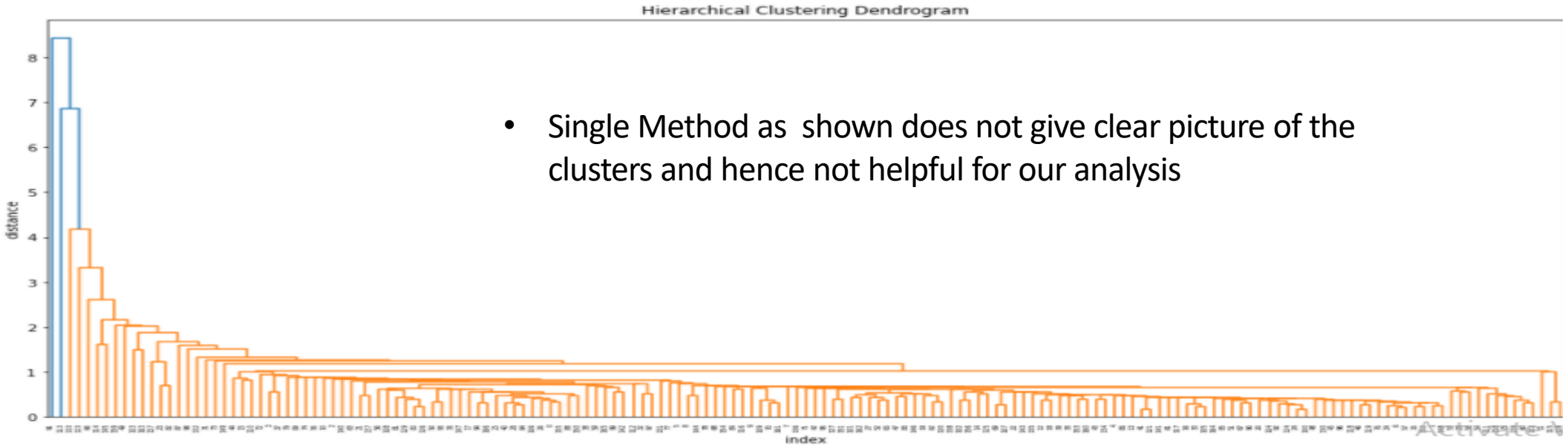
Visualizing the Data:



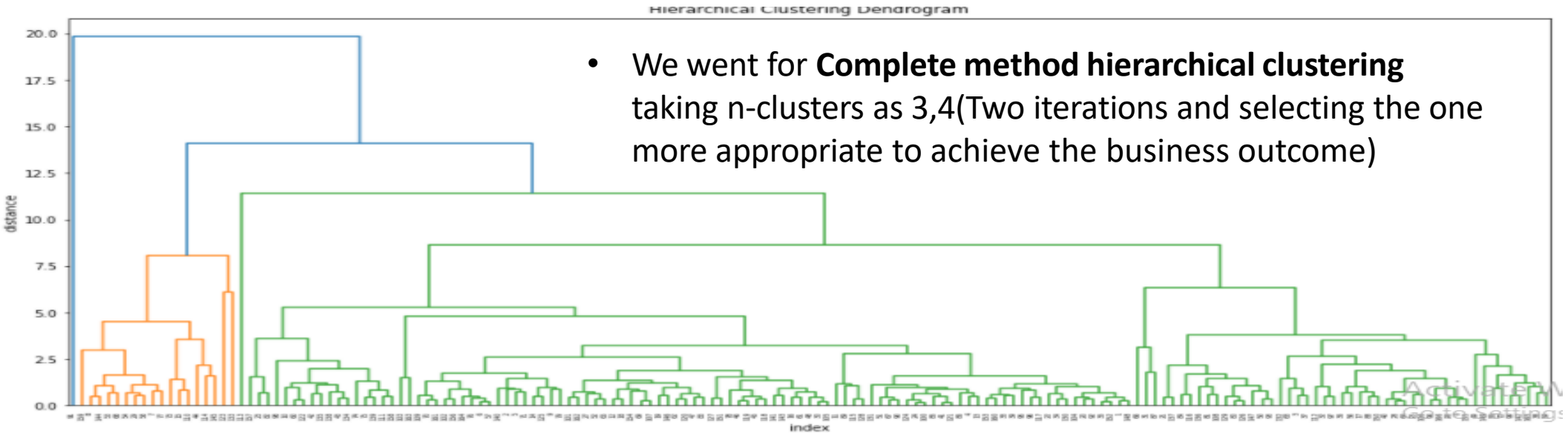
- Looking at the heatmap, we see that few variables like (total fertility, child mortality) , (income , gdpp) and (imports and exports) have high correlation.
- We did standardized scaling to standardize all parameters on cleaned, outlier removed data.

Hierarchical Clustering

Single hierarchical clustering



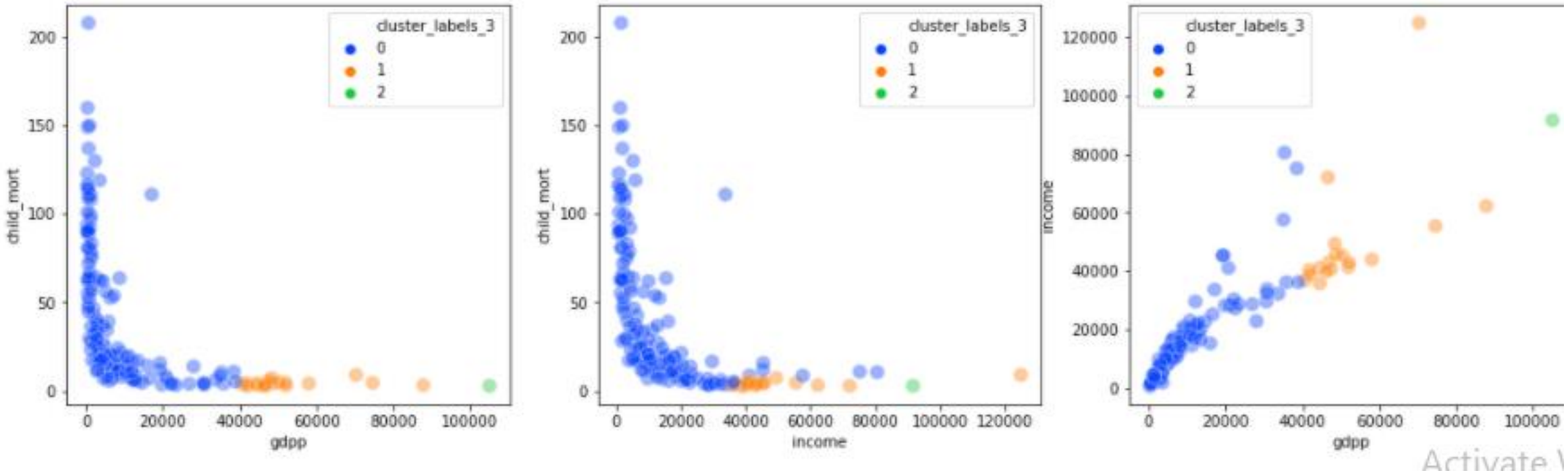
- Single Method as shown does not give clear picture of the clusters and hence not helpful for our analysis



- We went for **Complete method hierarchical clustering** taking n-clusters as 3,4(Two iterations and selecting the one more appropriate to achieve the business outcome)

Complete hierarchical clustering

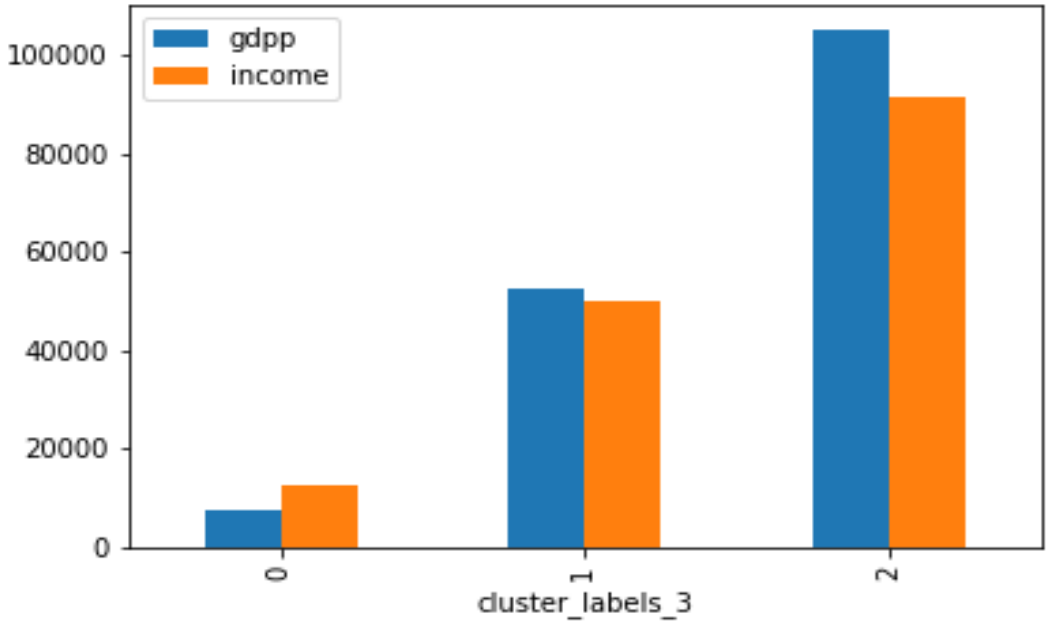
Hierarchical Clustering(k=3)

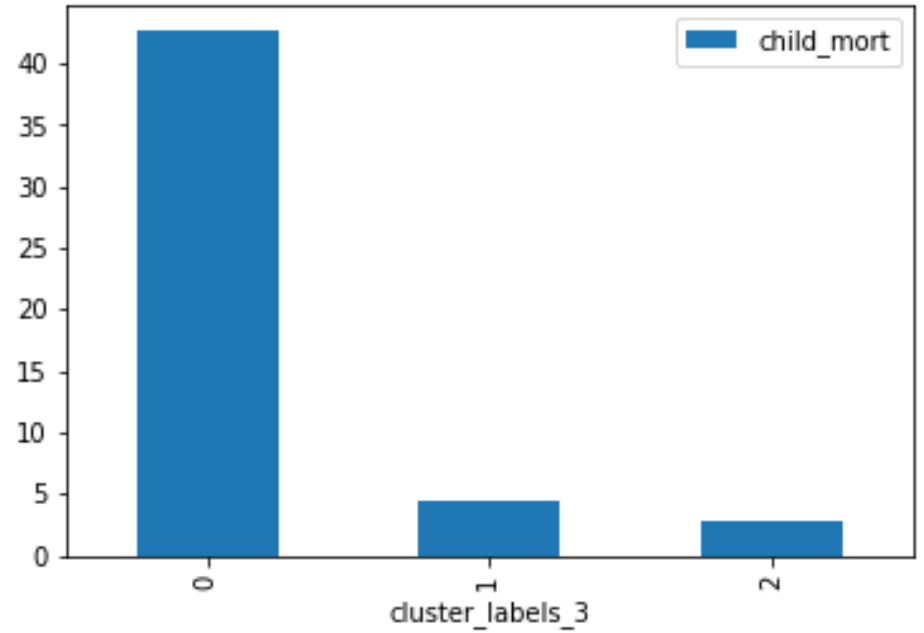


- Scatter plot of gdpp , income for various clusters. We see that for cluster 0 , both gdpp and net income per person are very low.

- Scatter plot of health spending , child mortality for various clusters. We see that for cluster 0, the health spending as % of gdp of few countries is lower and for those countries -the child mortality is very high.

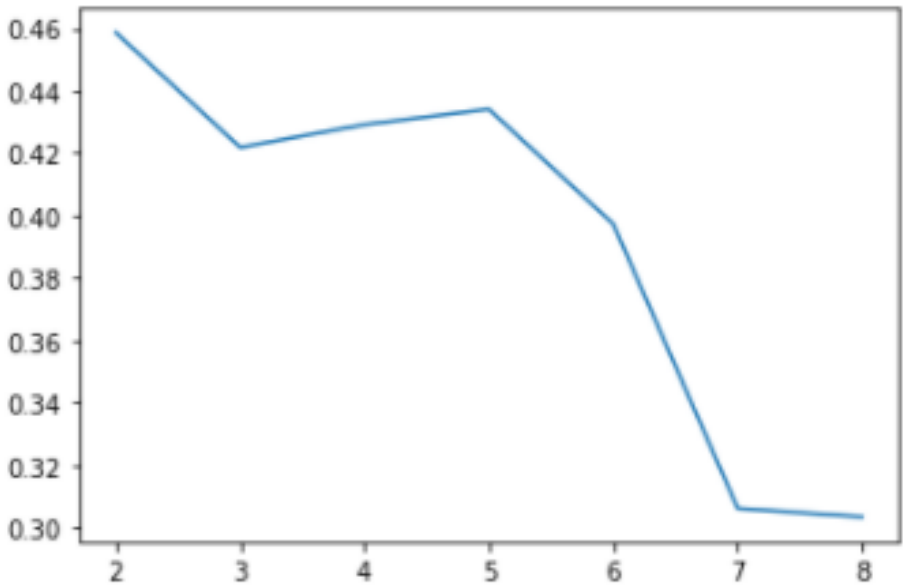
Interpretation from Hierarchical Clustering(k=3)



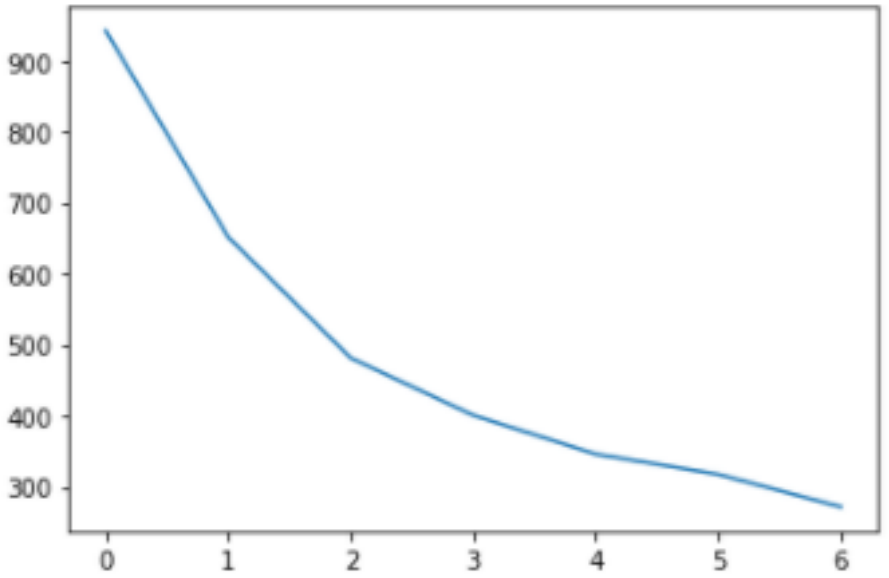


- Cluster 0 has the Highest average Child Mortality rate of **~42** when compared to other 2 clusters, and Lowest average GDPP & Income of **~ 7551 & 12641** respectively. All these figures clearly makes this cluster the best candidate for the financial aid from NGO.
- Cluster 0 comprises of **~89%** of overall data, and has ~148 observations in comparison to 167 total observations This seems to be a problem. This means that Hierarchical clustering is not giving us a good result as 89% of the data points are segmented into that cluster. We also saw that increasing the cluster number is not solving this problem. We will **perform K-Means Clustering** and check how that turns out to be

K-means Clustering



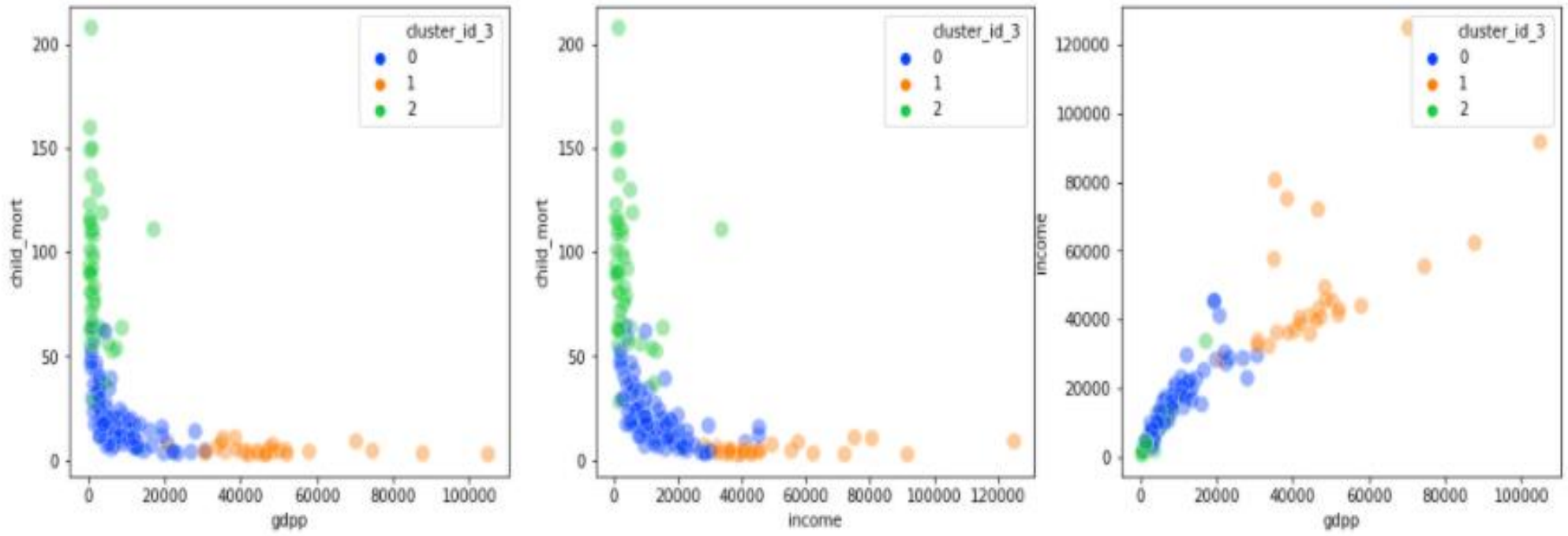
Silhouette Analysis



Sum of Squared Distances

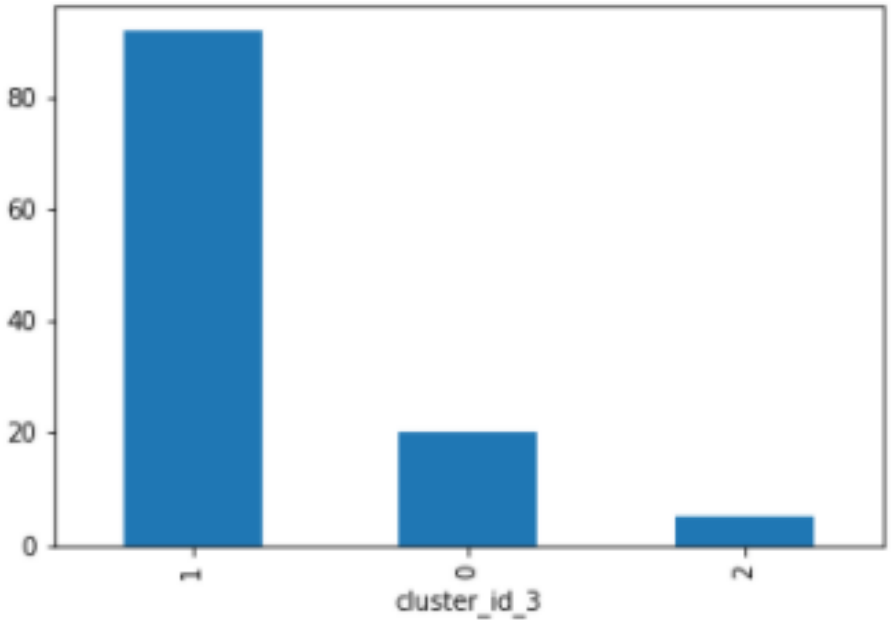
- From the above validations(Elbow Curve & silhouette analysis), we could see that 3,4 or 5 clusters are optimal number of clusters to be used. We will go ahead with Cluster=3 and 4 check which Cluster size is appropriate

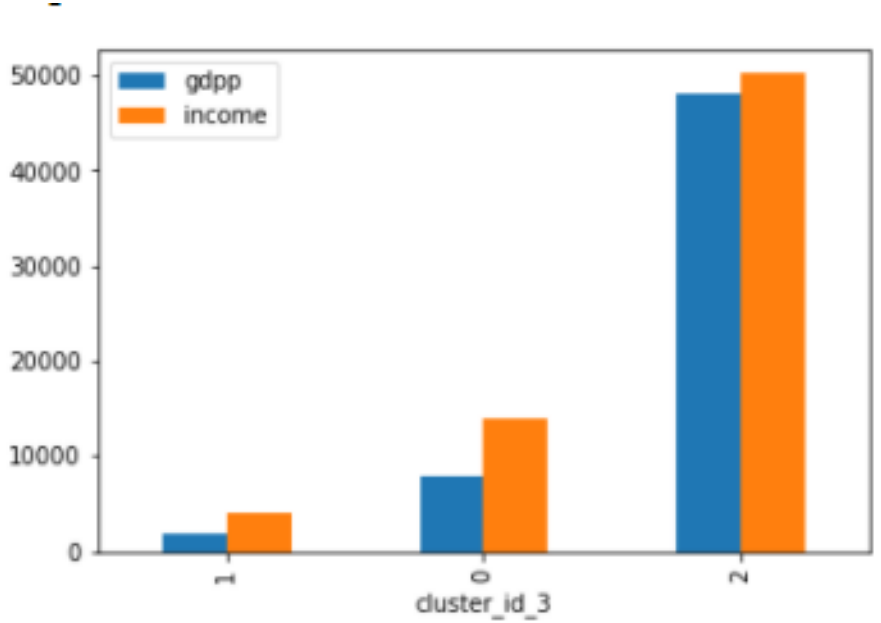
K-means Clustering(k=3)



- From above we can see Child mort is high for Cluster id 1 with respect to gdp and income and income, gdp is comparatively lower for this cluster. Cluster 1 seems to be countries which are under developed countries
- For Cluster id 2 we can see that the Child Mort is very low and income, gdp is on a very high side and seems that this cluster represents developed countries
- For Cluster id 1 we can see child mortality seems to be on a lower side and also income is higher but lower than income of Cluster id 2 and seems to be countries which are either developed/developing country

K-means Clustering(k=3)





- Cluster 1 has the Highest average Child Mortality rate of **~92** when compared to other 2 clusters, and Lowest average GDPP & Income of **~ 1909 & 3897** respectively. All these figures clearly makes this cluster **the best candidate** for the financial aid from NGO.
- We could also see that Cluster 2 comprises of **~29%** of overall data, and has ~48 observations in comparison to 167 total observations

Cluster Summary

The final model generated 3 clusters:

- Under-Developed Countries
  - Developing Countries
  - Developed Countries

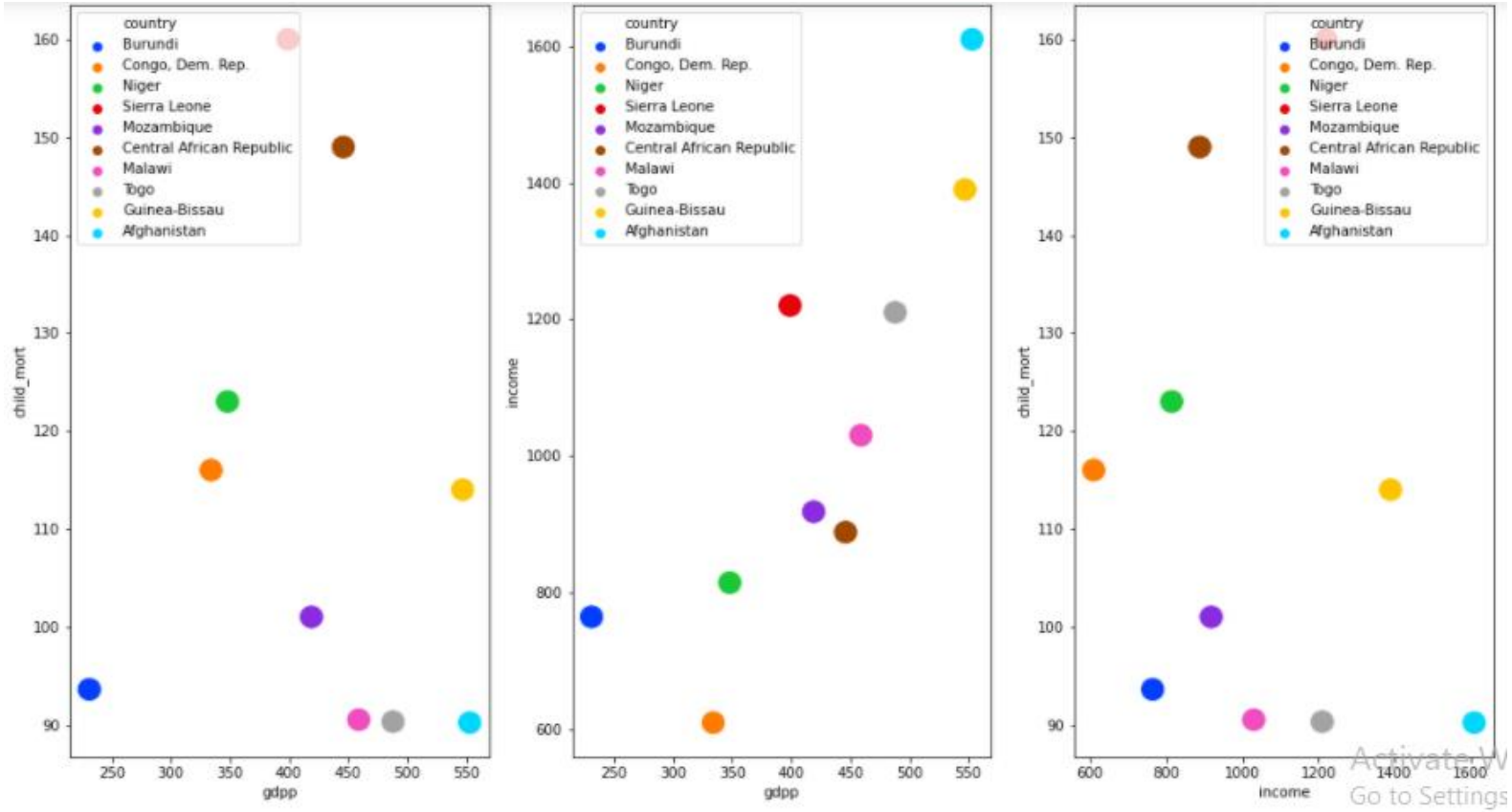
cluster_id_3	child_mort	income	gdpp	Observation
Under-Developed Countries	91.61	3897.35	1909.20	48
Developing Countries	20.36	13968.02	7979.91	91
Developed Countries	5.05	50178.57	48114.28	28

- We concluded the list of Top-10 UDC countries from our target Cluster('Under-Developed Countries') based upon median values of income, gdp and child mortality rate

List of Top 10 UDC

Countries where filtered based upon below factors:

- Lowest GDP
- Lowest Income
- Highest Child Mortality Rate



Country
Burundi
Congo, Dem. Rep.
Niger
Sierra Leone
Mozambique
Central African Republic
Malawi
Togo
Guinea-Bissau
Afghanistan