Assignment-based Subjective Questions

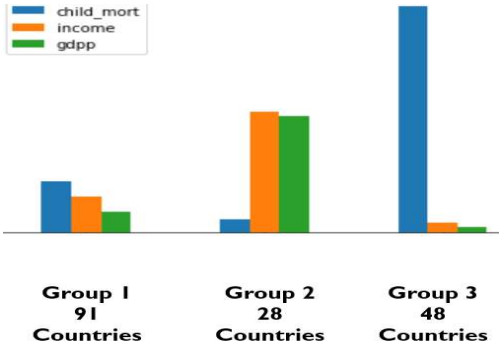
Question 1:

Assignment Summary

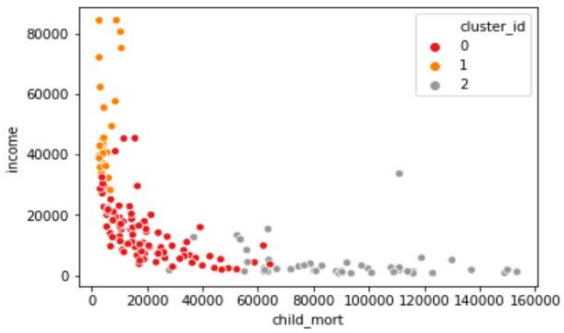
Problem is to categorise the countries using some socio-economic and health factors that determine the overall development of the country and suggest the countries which the CEO needs to focus on the most.

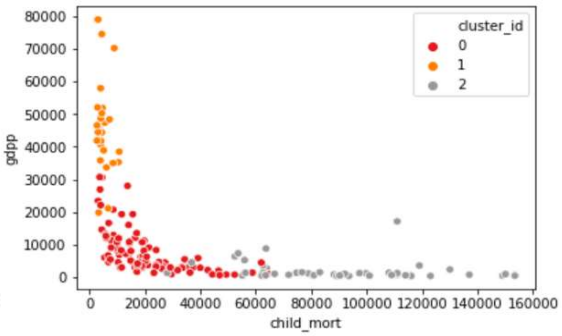
Process –

1. Data inspections and EDA for the data set – data cleaning / Univariate and Bivariate analysis
2. Outlier Analysis – to observe and remove outliers to reduce and capping performed for outliers beyond 1-99 percentile
3. K means and Hierarchical Clustering – both analyses performed, and similar results seen. Complete linkage considered for Hierarchical to obtain more stable clustering.
4. Analysis of the clusters and identification of group that is in dire need of aid by comparing how the 3 variables - [gdpp, child_mort and income] vary for each cluster to recognise and differentiate the clusters of developed countries from the under developed.



5. Visualisations on the clusters that have been formed





6. Results - K means and Hierarchical clustering process depend on previous analysis. But the outcome remains same after both types of analysis which indicates that Burundi, Liberia, Congo, Niger, Sierra Leone are the candidates for funding.

Question 2: Clustering

a) Compare and contrast K-means Clustering and Hierarchical Clustering.

Both are methods to identify similar groups of data in a dataset. The objective being performing grouping in a manner which results in each data points in group being like other entities of that group and different than those of other groups.

K means Clustering is an iterative clustering algorithm that aims to find the best local centroid for clusters until no further optimization is seem with iteration.
Hierarchical Clustering builds hierarchy of clusters. The algorithm starts with all the data points being assigned to a cluster of their own and subsequently 2 nearest clusters are merged together and it terminates in the final creation of 1 single cluster.

b) Briefly explain the steps of the K-means clustering algorithm.

K means clustering algorithm takes all the data points and groups them into K clusters.

The basic K-means algorithm has the following steps

1. Choosing K random initial centroid values

Assignment -

2. Allocating every data point to the closest centroid based on distance.

Optimization -

3. Calculate actual centroid values of the clusters formed / mean of all data points in a cluster after the previous assignment.
4. Assign data points to new/improved clusters based on distance i.e. update the cluster centroid associated to each data point
5. Keep on optimizing till the centroid location no longer updates i.e. we achieve optimal grouping.

c) How is the value of 'k' chosen in K-means clustering? Explain both the statistical as well as the business aspect of it.

Quantitative and Qualitative analysis is done for the selection of value of K where K is the number of clusters to perform the clustering. The K to be selected could be one of many choices and should be the value that gives the best tightness of clusters or the value that is most useful to the business scenario for the business problem being approached.

d) Explain the necessity for scaling/standardisation before performing Clustering.

Standardization / Data normalization / Feature scaling refers to process of rescaling the values of the variables in your data set so that they share a common scale. It is important where each variable has a different unit, or the scales of each variable vary from each other. When each variable means something different the fields could not be directly compared since, they are not equivalent or in the same range.

This is important as a pre-processing step before Clustering because groups are defined based on the distance between the points. If this is not handled the field with wider range of values may end up being the primary driver of the cluster. Standardisation helps avoid this by making relative weight of each variable equal by converting each variable to a unitless measure or relative distance.

e) Explain the different linkages used in Hierarchical Clustering.

Linkage is representative of the distance between 2 clusters.

Single Linkage – indicates the shortest distance between a pair of observations from 2 clusters. It could also point to observations in different clusters being closer to each other compared to observations in their own clusters. It implies that the clusters are made up of sparse and not close-knit observations.
Complete Linkage – indicates the largest distance between a pair of observations from 2 clusters / farthest pair of observations in 2 clusters. It implies that the clusters are very close-knit among themselves and are tighter clusters.
Average Linkage – indicates an average inter-cluster distance between observations. Average inter-cluster distance is found by adding distance between each pair of observations in each cluster divided by the number of pairs.
Complete Linkage and Average Linkage are the distance metrics that are more important measures since they imply clusters with proper Tree like structure in Hierarchical Clustering and moving from 1 cluster to n clusters (each observation in its own cluster), the clusters become more and more similar almost always.