**Executive Summary – Country Segmentation**

In this project, we are analyzing the principal components and clustering the countries into different clusters based on their socio-economic and health factors. The dataset used has one record for each country and the variables represent socio-economic factors of the country like GDP, income, child mortality rate, exports, imports, life expectancy, health, inflation, and total fertility rates. Analysis is done using the following unsupervised learning techniques:

* Principal Component Analysis (PCA) using prcomp() and principal()
* Hierarchical clustering using hclust
* Partitioning clustering using K-Means

On performing exploratory data analysis on the dataset using histograms, box plots, pair plots and correlation matrix the following observations are made:

* Dataset has no missing values (167 observation and 10 variables).
* Data is not normally distributed and is skewed.
* Some variables are in a totally different scale; hence data needs to be scaled.
* There are outliers on all numeric/integer variables.
* There are few variables with very high correlation like GDP & income, child mortality rate & life expectancy, etc.

The first unsupervised learning technique applied is Principal Component Analysis. This showed that 95% of variability in the dataset can be explained by the first 5 principal components and 63% of variability can be explained by the first 2 principal components. The variables that contributed the most to principal components are GDP, income, and life expectancy. The functions used for this analysis are prcomp(), and principal() using unrotated and rotated components. Various plots like scree plot, bar plot and network graph were plotted to show the contribution of variables to the principal components.

The next learning technique used on the dataset is hierarchical clustering using hclust function. Dendrograms are plotted and cluster size is analyzed with different distance metrics and clustering methods. The clusters formed using Canberra distance metric and ward.D clustering method gave equal sized clusters. The grouping of countries into three clusters looked good and relevant to this analysis.

The final unsupervised learning technique applied on the country dataset is partitioning clustering using k-means. Clustering is done on the dataset with different values for K. Looking at the clusters plotted using principal components and cluster size, it looked like K=3 performed well. Gap statistic methods shows the optimal number of clusters as 3. Silhouette plot also confirmed that the grouping has very less possible grouping errors.

Based on the above analysis, three is the number of clusters chosen. The clusters numbers are assigned to the dataset. Based on boxplots and means values of variables, clusters are named as “Developing Countries”, “Developed Countries” and “Under-Developed” countries. Thus, the country dataset is successfully categorized into different groups based on socio-economic status.