**STA6704 – Final Project**

**Country Segmentation – Unsupervised learning**

**Project Goal:**

The goal of this project is to apply unsupervised learning techniques on the country dataset and to cluster countries based on their socio-economic factors. This will help to determine the overall development on the country and identify the countries that needs more help for their development. The methods used for analysis in this project are:

* Principal Component Analysis (PCA) using
* prcomp()
* principal()
* Hierarchical clustering using hclust
* Euclidean, complete linkage
* Canberra, ward.D
* Partitioning clustering using K-means.

**Dataset:**

The dataset has ten variables which includes socio-economic and health factors such as child mortality rate, exports, health, imports, income, inflation, life expectancy, total fertility rates and GDP for each country. There are 167 countries in the dataset with one record per country. Below snapshot shows the number of variables, observations, and sample of the data in country dataset. It can be seen from the structure of the dataset that only the country column consists of characters and all the other variables are numeric/integers. There are no ordinal or categorical data in the dataset.

Text, letter

Description automatically generated

These are the column definition that can help in understanding the input variables.

Text

Description automatically generated

The summary of dataset shows the standard statistical measures like minimum, maximum, mean, median, etc for each of the variables.

A picture containing text

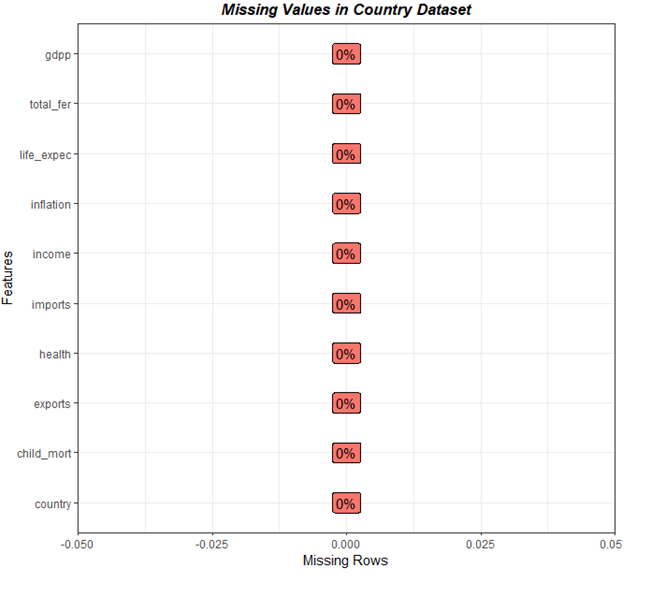
Description automatically generated

**Exploratory Data Analysis (EDA):**

To gain more insight of the dataset, we will perform some analysis.

1. **Missing Values:**

All the columns and rows have data. Hence, there are no missing values in any of the features in the dataset. Missing data plot is shown below.

****

1. **Univariable Analysis – Histogram:**

Histogram shows the frequency of values. Here, we are grouping each variable in the dataset and plotting a histogram. From the plots, we can say that most of the variables do not have a normal distribution.

Chart, histogram

Description automatically generated

1. **Univariable Analysis – Boxplot:**

Boxplot shows the statistical summary of the variables. The plot below shows boxplots of 7 variables from the dataset. Income and gdpp are excluded as the scales are quite different for these 2 variables from the others. This means we have to scale the data.

Chart, box and whisker chart

Description automatically generated

The plot below includes all the input variables from the dataset. It shows that all variables have outliers. Since there can be countries with extreme values like very high child mortality rate or very low life expectancy as compared to the rest of the world, we will not impute or remove these values. These datapoints which looks like outliers could even be the countries whose development is very less and is in dire need of help.

Chart, box and whisker chart

Description automatically generated

1. **Bivariate Analysis:**

Pair plot for all the input variables is shown below. The scatter plot for each pair of variables helps to see the relationship between them. The correlation coefficient is also displayed inside the scatter plot.

Some of the observations from the pairplot:

* total\_fer and child\_mort has strong positive relationship.
* Child\_mort and gdpp has negative relationship.
* Life\_expec and gdpp has positive relationship.
* total\_fer and life\_expec has negative relationship.

Diagram

Description automatically generated

1. **Correlation Matrix:**

Chart, bubble chart

Description automatically generated

Correlation matrix shown above displays the correlation coefficient between the variables. Bigger and darker circles represent stronger correlation. Higher correlation values on positive side are displayed in darker blue and higher correlation on negative side is displayed in darker red.

Some observations from the correlation matrix:

* Income and gdpp are highly correlated with correlation coefficient of 0.9. Higher the income of the country, higher its gdpp
* Child\_mort and tot\_fer are highly correlated with correlation coefficient of 0.85.
* Child\_mort and life\_expect have high negative correlation with correlation coefficient of -0.89. A country with high child mortality rate will have lesser life expectancy.

1. **Correlation Network graph:**

Another representation of correlation is shown in the graph below. Each variable is represented by a node. Closer the nodes and stronger the lines between them represents higher correlation between the two variables. Red represents negative correlation and green represents positive correlation.

Chart, radar chart

Description automatically generated

**Data is scaled and the following analysis are performed.**

**Unsupervised learning:**

**I. Principal Component Analysis (PCA):**

**(a) PCA using prcomp():**

The first unsupervised learning technique applied on the dataset is principal component analysis. Principal components using prcomp are shown in image below:

Text

Description automatically generated with medium confidence

The standard deviation, proportion of variance and cumulative variance of the principal components are shown below. About 95% of the variability in the dataset can be explained by the first 5 principal components.

Text

Description automatically generated

The eigen values and percentage of various of the principal components is shown below. The first dimension explains about 46% of variability and second dimension explains about 17% of variability in the dataset.

Text

Description automatically generated

The Scree plot displays the percentage of variance of each principal component.

Chart, histogram

Description automatically generated

The next two bar plot shows the contribution of each variable to principal components PC1 and PC2. The variables that contribute more to first principal component (PC1) are life\_expec, gdpp, income and exports. For PC2, the variables health and life\_expec contributes the most.

Chart, waterfall chart

Description automatically generated

Chart, waterfall chart

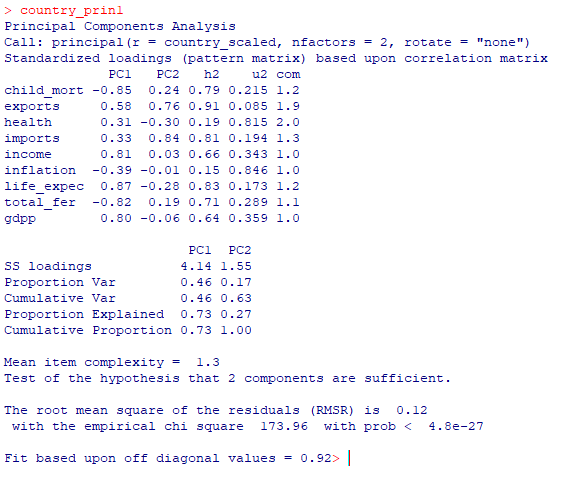
Description automatically generated

This plot also shows the contribution of variables to principal components PC1 and PC2.

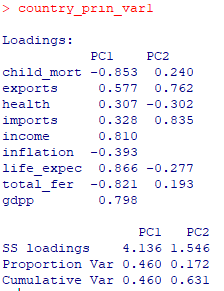
Chart, radar chart

Description automatically generated

**(b) PCA using principal():**

Principal component analysis is done on the same dataset using a different function principal(). Here we are looking at only first 2 principal components by setting the nfactors parameter as 2. Also, the rotate parameter is set to none. ****

Loadings of the principal components PC1 and PC2 shown below. The first two principal components explain 63% variability in the data.

****

The network graph shows the contribution of variables to principal components PC1 and PC2 without rotation of components. Each node(square) represents the variables and line/arc shows the contribution of the variables to the respective principal component PC1 and PC2 (specified as circles 1 and 2). Green colored arc represents positive and higher contribution. Red represents negative values. Stronger the line/arc, higher the values.

Chart, radar chart

Description automatically generated

From the network graph,

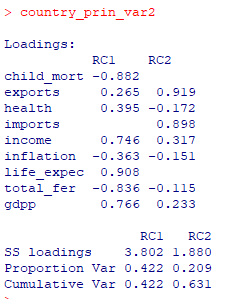
* life\_expec, gdpp and income contribute more to PC1.
* variables imports and exports contribute more to PC2.

Now, the principal component analysis is done using the same function principal() but with rotated components. This is done by setting a value for rotate function. The type used here is varimax. The contributions made by variables to the first two principal components are shown below:

Text

Description automatically generated

The loadings values for the principal components with rotated components.

****

Like the previous network graph, this also shows the variables contribution to principal components that are rotated. The graph is also similar, except the contribution for PC1 from exports is reduced here.

Chart, radar chart

Description automatically generated

**II. Hierarchical Clustering:**

The next method used for analysis the data is hierarchical clustering using hclust function.

1. **hclust (Euclidean, complete):**

The dendrogram of hierarchical cluster formed using default Euclidean distance metric and complete linkage clustering method is shown below.

A picture containing chart

Description automatically generated

This dendrogram shows the tree using different colors for each cluster if they are cut for three clusters.

Chart, box and whisker chart

Description automatically generated

This is the scatter plot of three clusters using the principal components PC1 and PC2.

Chart, scatter chart

Description automatically generated

From the above plot, we can say that size of cluster three is very small. Also looking at the country’s name, we can say that they need not belong a separate cluster. The clusters formed by hierarchical clustering complete linkage method did not create equal sized clusters. Hence, we can try to create equal sized clusters using a different distance metric and clustering method.

1. **hclust (Canberra, ward.D):**

Hierarchical clustering is performed on the dataset using Canberra distance metric and ward.D clustering method. The following dendrograms shows the clusters created.

Graphical user interface, diagram

Description automatically generated

The colored dendrogram clearly shows that the clusters are more equal sized as compared to the previous method.

Chart

Description automatically generated

This is the scatter plot of three clusters using the principal components PC1 and PC2.

Chart, scatter chart

Description automatically generated

From the above plot, we can see that the cluster sizes are equal. So, this method seems to have created good clusters. Cluster 1 marked in red has countries like Haiti, Nigeria. Congo, Rep and other African countries which might represent under-developed countries. Cluster 2 marked in green includes countries like Brazil, Malaysia, and some Asian and south American countries, which might represent developing countries. Cluster 3 marked in blue has countries like United States, Australia, Netherlands which might represent developed nations. Hence from this hierarchical clustering, we can see that cluster 1 could potentially be the list of countries that is not developed socially or economically and might need help.

**III) K-Means Clustering:**

The last technique to be applied on the dataset is k-means clustering. We will try to create various number of clusters and try to find best value for K and analyze the countries assigned to the clusters. Using different values for k (k = 2, 3, 4 and 5) clustering has been done and plotted against the principal components PC1 and PC2.

**K= 2 Cluster Plot**

Chart, map

Description automatically generated

**K = 3 Cluster Plot**

Chart

Description automatically generated

**K = 4 Cluster Plot**

Map

Description automatically generated

**K = 5 Cluster Plot**

Map

Description automatically generated

From the above cluster plot for various values of K, 2 and 3 seems to be good options in terms of cluster size. However, for K=2, developed countries like United States belong the same cluster as developing nations like Thailand, Brazil, etc. Also, the number of datapoints on some clusters for K = 4 and K=5 is very less. Hence, it looks like 3 clusters will be a better choice.

To find the optimal number of clusters, the following plots are used:

* **Elbow Method:**

Using the elbow method, the optimal number of clusters can be 4 or 6.

Chart, line chart

Description automatically generated

* **Silhouette Method:**

Using this method, optimal number of clusters is suggested as 5.

Chart, line chart

Description automatically generated

* **GAP statistics Method:**

As per plot using gap statistic method, optimal number of clusters is 3.

Chart, line chart

Description automatically generated

Different methods suggest different optimal number of clusters. Let see the silhouette plot for all the clusters. Silhouette plot for different values of K shown below.

Chart, funnel chart, surface chart

Description automatically generated

Confidence of the data points being grouped into clusters looks good for all values of K. But as mentioned earlier, the cluster size of one of the clusters is only 3 for k = 4. Similarly, for K = 5, there are two clusters with size 1 and 3.

For K=2, developed countries like United States belong the same cluster as developing nations like Thailand, Brazil, etc. This grouping of certain countries for K = 2 is not correct from the project goal perspective. Hence, we can decide the number of optimal clusters to be 3 for this analysis.

**Merge cluster value:**

After deciding the number of clusters as 3, we will assign the corresponding cluster number to each observation in the original dataset. Here a new column kmeans3 is added to the dataset which represents the cluster number to which the observation belongs.

Table

Description automatically generated

Let’s analyze the cluster assignment with some plots to identify the characteristics of the countries assigned to that cluster.

Boxplot of all variables grouped by clusters is shown in the image below

Diagram

Description automatically generated

Some of the observation from the boxplot clusters

* Cluster 1: Has high child\_mort, low income, low life\_expec and low gdpp
* Cluster 2: Has moderate values for most variables
* Cluster 3: Has low child\_mort, high income, high life\_expec and high gdpp

Mean values of variables based on clusters:

Text

Description automatically generated

From the boxplots and average values for all variables as shown above, we can say that:

* Cluster 1 => Under-Developed Countries
* Cluster 2 => Developing Countries
* Cluster 3 => Developed Countries

Adding country category to the dataset:

Table

Description automatically generated

**Visualizing variables – Bar plot:**

In the following bar plots, sorted variable values are plotted by country and color is assigned by cluster category determined by kmeans clustering. We will consider 4 variables which has clear distinction in the values between the clusters in the group box plot namely, child\_mort, gdpp, income and life\_expec.

Child mortality is higher in under-developed countries followed by developing countries and it is least in developed countries

Chart, histogram

Description automatically generated

GDP per capita is higher in developed countries followed by developing countries and it is least in under-developed countries

Chart, histogram

Description automatically generated

Income is similar to GDP per capita, higher the value, more developed a country is.

Chart

Description automatically generated

Life expectancy is also higher for developed nations and decreasing towards developing countries and least values for under-developed countries.

Chart

Description automatically generated

**Visualizing variables – Scatter plot:**

Like bar plots above, we will plot the selected 4 variables on scatter plot. As mentioned earlier, child\_mort and life\_expec have very high negative correlation whereas income and gdpp have very high positive correlation. Let’s view the clusters in the scatter plot.

Chart, scatter chart

Description automatically generated

**List of countries in each cluster/category:**

A picture containing table

Description automatically generated

Table

Description automatically generated

Table

Description automatically generated

**Visualizing the country clusters on map:**

Map

Description automatically generated

From the above map, we can say that cluster 1 (Under-Developed) has mostly African and some Asian countries. Cluster 2 (Developing) includes mostly South American and Asian countries. Cluster 3 (Developed) includes countries from Europe, North America, and Australia. This supports the clustering done based on socio economic factors. (Note: Gray color on the map refers to countries that does not have data in the original dataset)

**Conclusion:**

Various unsupervised learning techniques like principal component analysis, hierarchical clustering and partitioning clustering have been applied to the dataset. The countries are categorized into 3 clusters, that is developed, developing and under-developed categories and multiple visualizations are created to view the data and clusters. This segmentation can be used to suggest the set of countries that might be in dire need of help. Thus, country segmentation has been done successfully.