Question 1: Assignment Summary

Briefly describe the "Clustering of Countries" assignment that you just completed within 200-300 words. Mention the problem statement and the solution methodology that you followed to arrive at the final list of countries. Explain your main choices briefly(why you took that many numbers of principal components, which type of Clustering produced a better result and so on)

Ans: The task is to cluster the countries by the factors given below to filter out list of 5 countries in direst need of aid.

Column Name	Description
country	Name of the country
child mort	Death of children under 5 years of age per 1000 live births
exports	Exports of goods and services. Given as %age of the Total GDP
health	Total health spending as %age of Total GDP
imports	Imports of goods and services. Given as %age of the Total GDP
Income	Net income per person
Inflation	The measurement of the annual growth rate of the Total GDP
life_expec	The average number of years a new born child would live if the current mortality patterns are to remain the same
total_fer	The number of children that would be born to each woman if the current age-fertility rates remain the same.
gdpp	The GDP per capita. Calculated as the Total GDP divided by the total population.

Solution Methodology:

Part A: Outlier treatment

Plotting all the 9 factors and looking for outliers keeping in mind that we could not remove outliers lower on scale for factors gdpp, income because lower income and gdpp are the focus of this study.

Part B:

PCA: Further to reduce the dimensionality of data PCA was applied and only 5 components were retained which explained 95% variance in data.

Part C:

Using Elbow method and Silhouette Score to decide K= number of clusters and applying K Means clustering on number of clusters = 10

Part D:

Analysing the result of clusters low on life expectancy, gdpp, income and high on child mortality from the box plots plotted for all the factors for each cluster.

Part E:

Clustering using hierarchical clustering method for single and complete linkage and cutting tree to get number of clusters = 15 based on tree structure and since looking or cluster of 5 countries.

Conclusion:

Analysing the result of hierarchical clustering for clusters low on life expectancy, gdpp and high on child mortality and select 5 countries using some subjectivity like stability and severity of situation and even comparing the result with K Means clustering.

- Sierra Leone
- Chad
- Central African Republic
- Burkina Faso
- Malawi

Question 2 **a:** Compare and contrast K-means Clustering and Hierarchical Clustering.

Ans:

I) Hierarchical clustering requires more computer resource than K Means clustering hence not suitable of big data size when we have lesser compute resource.

II) K Means clustering requires prior idea of clusters but hierarchical clustering does not

III) K Means work well with cluster having circle or sphere like shape.

Question 2 b:

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

Ans: Steps of K Means clustering

It randomly chooses data points as cluster centres.

Reassign each point to closest cluster centre based on Euclidean distance.

Recomputes the centroids

Does reassigning again and recompute centroids till no different centroid is found.

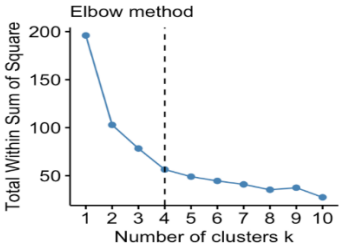
Question2 c) How is the value of 'k' chosen in K-means clustering? Explain both the statistical as well as the business aspect of it.

Ans:

It is chosen based on elbow method or silhouette score or business understanding of problem.

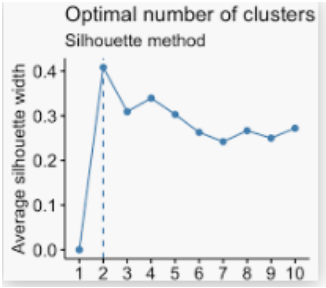
Elbow method: Total within-cluster sum of square (WSS) is minimized for chosen number of clusters.

For each K we plot the WSS and where there is bend in graph is taken as right choice or cluster.



Silhouette Score method:

It measures the how well each point lies within the cluster. For given number of Ks it calculate the silhouette score and selects the K for which score is maximum.



Business Understanding: Based on business understanding/need we could also select number of clusters since we could not take action for each cluster if number of clusters are high of distinguished from one another.

Question 2d) Explain the necessity for scaling/standardisation before performing Clustering.

Ans:

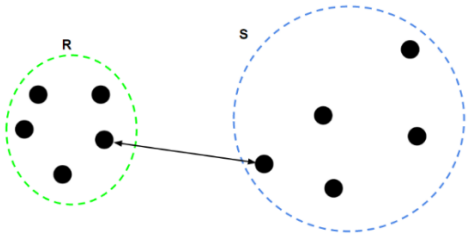
We need to scale the data to allow the algorithm not to get biased towards higher values present in features. For example, if we cluster based on latitude, longitude and price of the property as three features to cluster the similar properties. Without scaling the higher values in price column will bias the centroid towards itself.

Question 2e) Explain the different linkages used in Hierarchical Clustering.

Ans:

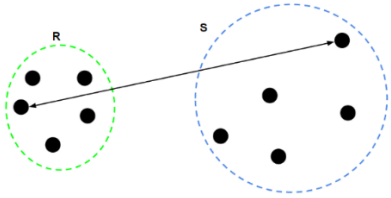
Single linkage:

It returns the minimum distance between two points in cluster p and q



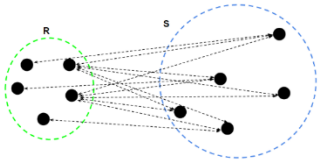
Complete Linkage

It returns the maximum distance between two points in cluster p and q



Average linkage:

It returns the arithmetic mean of distances between pair of points in cluster p and q.



Question 3a

Give at least three applications of using PCA.

Ans: PCA is used to reduce the dimension of data especially the image data with many features even correlated ones.

It is used for visualization by compressing the data into two components and use the scatter plot to visualise them earlier not possible for data having more than 2 features.

PCA is used to speed up the execution of machine learning algorithm by reducing the number of features and removing multi collinearity.

Question 3b) Briefly discuss the 2 important building blocks of PCA - Basis transformation and variance as information.

Ans:

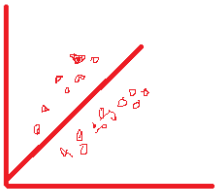
The PCA starts with covariance matrix to find the basis vectors orthogonal to each other so that features are uncorrelated in the new vector space while keeping the maximum information with reduced dimension.

Question 3c: State at least three shortcomings of using Principal Component Analysis.

Ans:

i)Components or features obtained as the result of PCA cannot be explained since are linear combination of original feature in different vector space / coordinate system. We could not explain a model to business user in terms of PCA features.

ii) PCA is not useful if original shape of data is in below shape since it finds the direction of maximum variance wrongly.



iii)It finds the basis vectors as the linear combination of standard basis but when it is not linear it will fail to correctly express it as a linear combination.