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)
Note: You don't have to include any images, equations or graphs for this question. Just text should be enough.

Answer:
According to the problem, CEO of the NGO wants to use the money they have collected strategically to provide help and support to the people of the backward countries.
We have to find the countries which require the direct aid from the NGO by using socio-economic and health factors that determine the overall development of the country.
For that case, we have performed the machine learning technique where we are about to find the countries by dividing them based on the factors such as income, gdpp, health and child mortality rate, etc.
The algorithm requires here is **unsupervised clustering** as there is no prior label provided in the dataset. But before that we have to find the features which are actually affecting the growth and development of the country.
So steps required finding the under developed countries are:

1. Finding the important features using **Principal Component Analysis**
    - For that reason I took 2 principal components using **scree plot**
2. With these features, perform **K-Means** clustering
    - First find the optimum number of clusters (k) needed using **elbow-curve** and **silhouette analysis**
    - Then performing k-means algorithm which divide the features into k clusters
    - Then finding the countries which have low gdpp, income, health and high child_mort
3. With the features we got after PCA, perform **Hierarchical clustering**
    - For finding the optimum number of clusters, perform single linkage and complete linkage algorithm
    - Find the optimum clusters by analysing the **dendrograms**
    - Divide features into clusters and the find the countries which are in direct aid from the NGO

I found same countries from K-Means and Hierarchical clustering algorithm which are:

- **Haiti**
- **Sierra Leone**
- **Chad**
- **Central African Republic**
- **Mali**

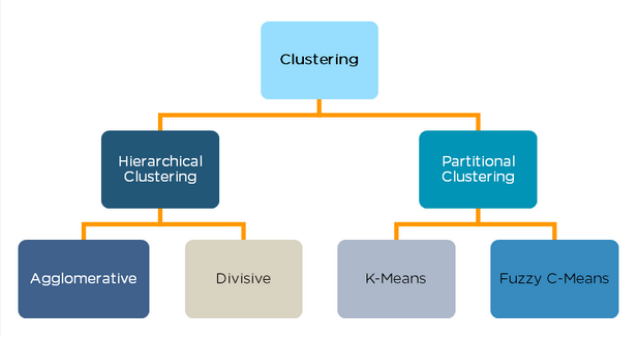
It’s easy to work with Hierarchical clustering as just by looking at the dendrogram we can find the clusters required. It does not require any prior knowledge of k.

Question 2: Clustering

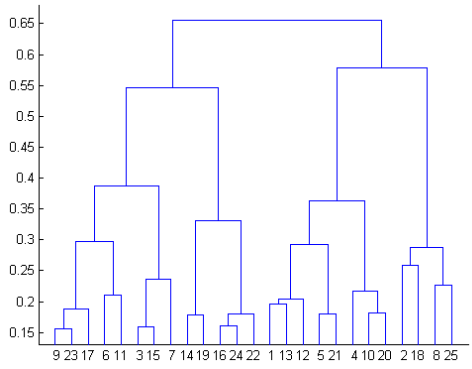
a) Compare and contrast K-means Clustering and Hierarchical Clustering.
b) Briefly explain the steps of the K-means clustering algorithm.
c) How is the value of 'k' chosen in K-means clustering? Explain both the statistical as well as the business aspect of it.
d) Explain the necessity for scaling/standardisation before performing Clustering.
e) Explain the different linkages used in Hierarchical Clustering.

Answers:

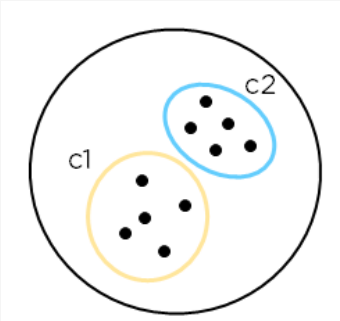
a) **Compare and contrast between K-means and Hierarchical clustering** :



- **K-means** are used for big data whereas **Hierarchical** clustering cannot handle big data. This is because the time complexity of K-means clustering is **linear** in nature i.e., O(n) and for Hierarchical clustering its **quadratic** i.e., $O(n^2)$
- **K-means** works on random selection of clusters therefore result differ every time we perform k-means algorithm while we can reproduce result in **Hierarchical** clustering.
- **K-means** works well when cluster is distributed like circle in 2-D and sphere in 3-D i.e., shape is hyper spherical in nature.
- **K-means** requires the prior knowledge of k i.e., we need to find the optimal value of k to divide the data points into k clusters. While in case of **Hierarchical** clustering we can find the number of clusters by interpreting the dendrogram.
- **K-means** does not consume much RAM as it happens in iteration while **Hierarchical** clustering requires huge amount of memory as it holds each data points.



Hierarchical clustering



K-Means clustering

b) **The steps of the K-means clustering algorithm**:

1. Start by choosing how many clusters we need by defining the value of K, this will be the initial cluster centres
   Let's say we have 10 data points and we want to divide in 2 clusters. Therefore the value of k = 2

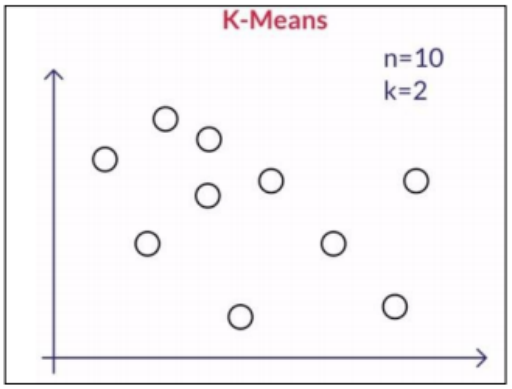


Fig 2: A set of 10 points to be divided into 2 clusters

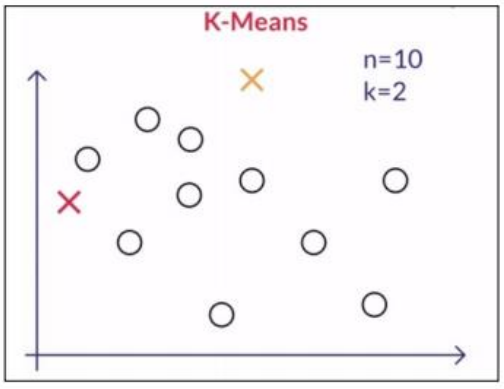


Fig 3: Choosing K random initial cluster centres

2. Assign each data point to its nearest initial cluster centres by measuring the distance between each data point to the cluster centres by calculating Euclidean distance between them.

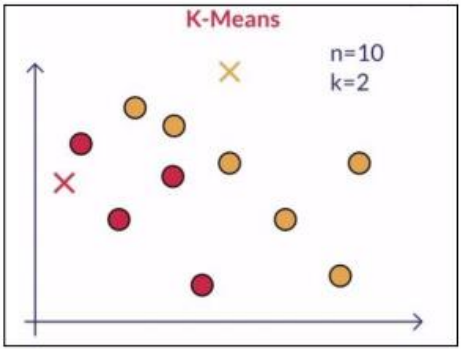


Fig 4: Assigning each data point to their nearest cluster centre

3. For each cluster, find the new cluster centres by calculating mean of all the data point in that cluster
4. Now again calculate the Euclidean distance of the data points with the newly selected cluster centres and reassign these data points with its nearest cluster centres.

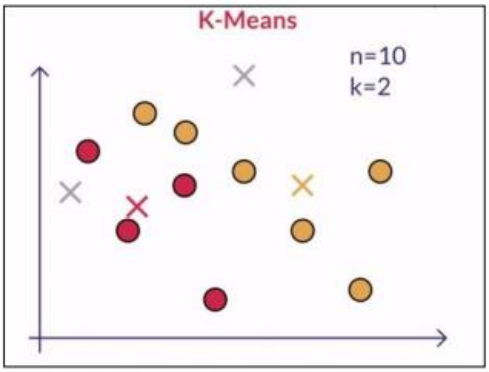


Fig 5: Updating the cluster centres

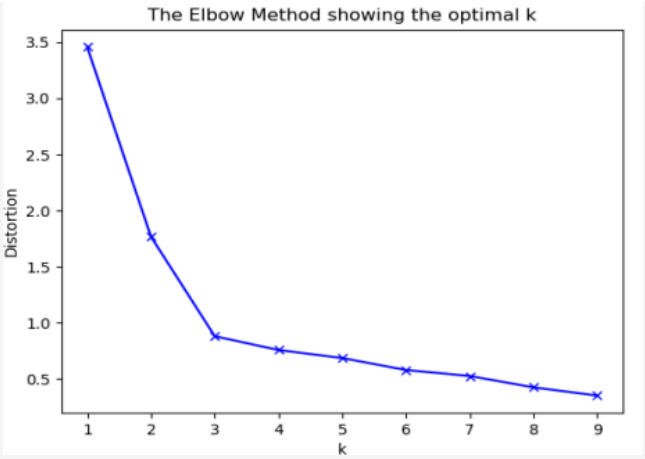
5. Repeat the steps 3 and 4 until no new cluster centres can be assigned. And finally we get the optimal clusters.

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

There are various methods to choose the value of k:
1. Elbow method
2. Average silhouette method

1. <u>Elbow method:</u>

- Calculate clustering algorithm for various values of k
- For each value of k, calculate the within-cluster sum of square (wss)
- Plot the curve of wss for every value of k
- The location of bend (knee) is generally considered as the appropriate value of k



2. **<u>Average silhouette method:</u>**

- Calculate clustering algorithm for various values of k
- For each value of k, calculate average silhouette score
- Plot the curve of silhouette score for different of values of k
- The location of maximum is considered as the appropriate value of k

Statistically, we select the optimum cluster where there is first bend in the elbow-curve and where average silhouette score is maximum.
But, the optimum selection of k using elbow-method and silhouette analysis depends on the business problem we are working on.

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

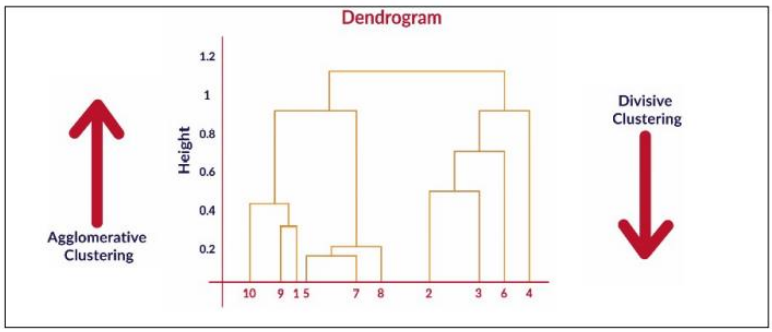
<u>Standardization/scaling</u> is the process of converting the data into z-scores with mean = 0 and standard deviation = 1
<u>Two reasons for scaling the data are:</u>

- Scaling down the data to same normal level avoids the large range of values to out-weight the small range of values while calculating the Euclidean distance.
- The data available might have different units. Standardization helps in making the attributes unit-less and uniform.

e) **Explain the different linkages used in Hierarchical Clustering.**

Linkage is the measure of dissimilarity between clusters having multiple observations.
Height of dendrogram represents similarity in the clusters.
There are 2 ways of interpreting the dendrogram:

1. Agglomerative
2. Divisive

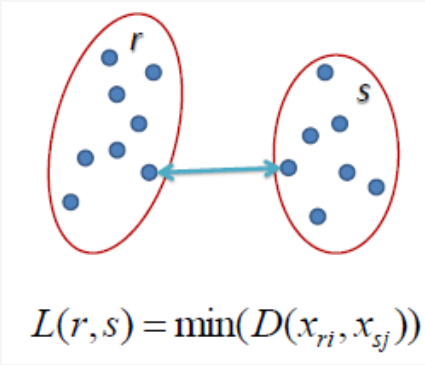


Different types of linkages are:

1. Single linkage
2. Complete linkage
3. Average linkage

1. **Single linkage:**

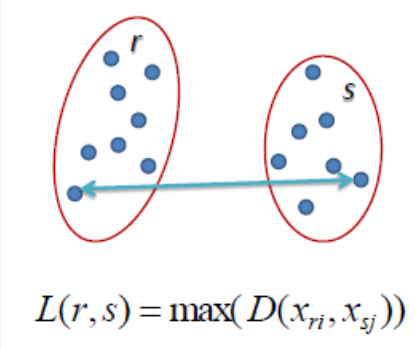
Distance between 2 clusters is defined as the shortest distance between points in the two clusters.



Where, ‘r’ and ‘s’ are two clusters

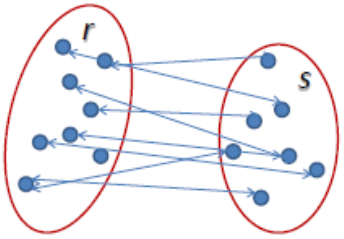
2. **Complete linkage:**

Distance between 2 clusters is defined as the maximum distance between any 2 points in the clusters.



3. **Average linkage:**

Distance between 2 clusters is defined as the average distance between any points of one cluster to every other point of the other cluster.



$$L(r,s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} D(x_{ri}, x_{sj})$$

Question 3: Principal Component Analysis

a) Give at least three applications of using PCA.
b) Briefly discuss the 2 important building blocks of PCA - Basis transformation and variance as information.
c) State at least three shortcomings of using Principal Component Analysis.

Answers:

a) Give at least three applications of using PCA.

Different applications of PCA are:

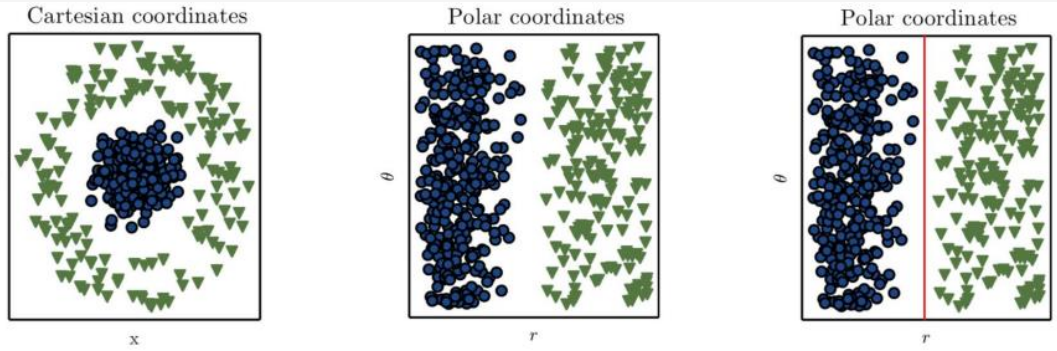
1. **Dimensionality reduction** – reducing variable by selecting important features from the data set
2. **Data visualization and EDA –** pairwise plot visualization are difficult when there are more variables and here PCA plays an important role by reducing dimensionality.
3. **Create uncorrelated features/variables that can be an input to a prediction model –** create features which are uncorrelated to each other
4. **Uncovering latent variables/themes/concepts –** PCA finds hidden features for the data
5. **Noise reduction in the data set –** it is mostly used for image processing

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

1. **Building block of PCA - Basis transformation:**

Sometimes the X and Y coordinates need to be represented based on angle and distance. For measuring the angle and distance of the coordinate we use **polar coordinate system**.

Basis transformation is the process of converting the information from one set of basis to another often for convenience and efficiency.



From the above figures we can see that for **Cartesian** coordinates, the separation is difficult but for **Polar** coordinates, the separation is easy.

Examples where basis transformation can be used are:

- Mechanical motion equations are significantly simplified and neat in polar coordinates.
- Electric field calculations are much cleaner in spherical coordinates.
- Dimensionality reduction: 3D world is captured and represented on a 2D screen.
- Dropping columns is basis reduction.

2. Building block of PCA - Variance:

- The more variance a column has, the more information that column or variable has for our data set.
- If two variables are correlated then drop one of them which add less value to the dataset which shows less variance.
- If variance of all the columns is distributed in the same way then we change the basis vector of the columns and drop the column with less variance.
- PCA helps in finding the best possible set of basis vectors for a given dataset in such a way that the variance is non-uniformly distributed among them – some columns now explain far more variance than the other columns. This makes it easier to choose which columns to keep and which to discard.

c) State at least three shortcomings of using Principal Component Analysis

1. The PCs have to be linear combinations of the original columns – for non-linear combinations we can use **t-SNE** as an alternate but it's quite expensive
2. PCA requires PCs to be uncorrelated/orthogonal/perpendicular – **ICA** (Independent Component Analysis) overcomes this drawback but is most of the time slower than PCA

3. PCA assumes low variance components are not very useful – this lead to loss of valuable information.