Question 1
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 -1

Problem Statement:
This assignment was aimed at giving recommendations for countries to Help International that are in dire need of funds. HELP International is an international humanitarian NGO that is committed to fighting poverty and providing the people of backward countries with basic amenities and relief during the time of disasters and natural calamities. It runs a lot of operational projects from time to time along with advocacy drives to raise awareness as well as for funding purposes.

After the recent funding programs, they have been able to raise around $ 10 million. Now the CEO of the NGO needs to decide how to use this money strategically and effectively. The significant issues that come while making this decision are mostly related to choosing the countries that are in the direst need of aid.

Data Set Summary
We were given a list of 167 countries together with their socio economic and health statistics and were asked to analyse these statistics and give recommendations on the list of countries based on our analysis.

Solution methodology:
The analysis first required a careful examination of data set. This included data understanding and cleaning. Data understanding include understanding size and shape of data, the various socio and economic factors etc. This was followed by data Cleaning which included checking for missing values, corrupt data, unwanted rows and columns, duplicate rows and outliers. We removed outliers

Then we performed Principal component Analysis (PCA). PCA preprocessing required first bringing the variables to common scale as PCA is sensitive to scale of data. So, we normalized data using Standard Scalar.

Then we did PCA to get Principal components. Then we did some analysis on the loading factor and correlation of Principal components obtained.

Then we formed a new data frame of components. The number of components in this data frame decided by drawing a scree plot. Scree is a plot between no of components and cumulative variance. We chose the no of components as 4 as it showed that with 4 components we could capture around 85 % of variance and reduce dimensionality to half.

Then we did Hopkin Statistic to see if proper clusters can be formed from data . We got a value greater than 0.7 which meant it is a data that is not random and supports clustering.

Then we proceeded to K Means clustering. We found value of K by applying Elbow Curve Method first and then Silhouette Coefficient Analysis to get the value of K which came out to be 3. This also has business sense as we aim to divide the countries into three groups under developed, developing and developed. We ran the KMeans of 3 clusters by providing it a fixed initial salt so that the cluster number always remains constant despite repeated runs of algo.

On running the K-Means algorithm we got the cluster ids of each row. Then we plotted the cluster id against the original variables to understand what each cluster represents. We found that cluster 2 represented the underdeveloped set of nations as the original variables values highlighted poor economical factors lke gdpp, chld mortality, life expectancy and net income. These countries included Afghanistan, Benin, Burundi, Burkina Faso, Central African Republic, Chad, Congo Republic, Congo Democratic Republic etc.

Then we took the new data frame of components obtained after PCA and carried out Hierarchical clustering plotting the dendrogram. The business justification says we need to break into 3 well formed popular clusters representing under developed, developing and developed countries. but on observing the linkage we find that we get three well defined populated clusters in the dendrogram only when we take at least 5 clusters because two of these 5 clusters captures outlier countries and have 1 and 4 countries. so it makes sense to take 5 as no of clusters for Hierachical Clustering.

We found that cluster 0 represented the underdeveloped set of nations as the original variables values highlighted poor economical factors like gdpp, child mortality, life expectancy and net income. These countries included Afghanistan, Benin, Burkina Faso, Botswana, Burundi, Central African Republic, Chad, Congo Democratic Republic, Kenya etc.

After looking at both the cluster 2 of KMeans cluster and cluster 0 of Hierarchical Clustering we came up with a list of 7 countries that are in dire need of aid. This was done on the basis of their socio-economic statistics. The final list of countries that are recommend after analyses is:- **Afghanistan, Central African Republic, Burundi, Congo Democratic Republic, Haiti, Sierra Leone, Niger**

In terms of separation of cluster K-Means created clusters which were clearly separable and so did hierarchical clustering .

In terms of prerequisites, since K Means required calculation of K first it was an additional work required for K Means which was not required in Hierarchical clustering.

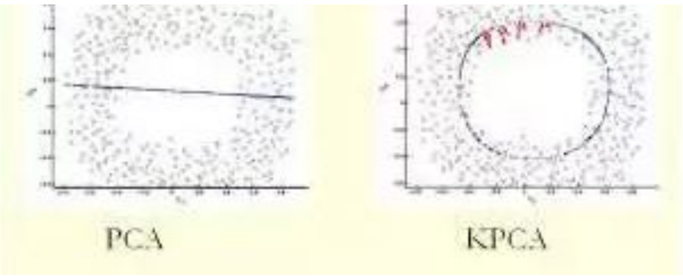
Question 2
State at least three shortcomings of using Principal Component Analysis.

Answer -2

The various Shortcomings of PCA are:-

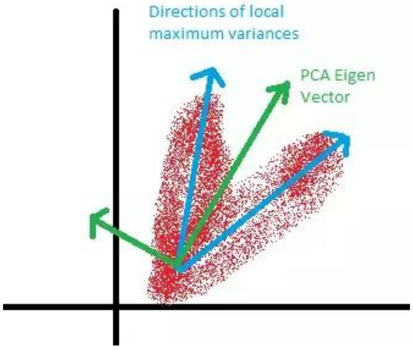
1. The standard PCA always finds linear principal components to represent the data in lower dimension. It is used basically before building linear models like

linear regression and logistic regression. Sometime, we need non-linear principal components.



If we apply standard PCA for the above data, it will fail to find good representative direction. Kernel PCA (KPCA) rectifies this limitation.

2. PCA always finds orthogonal principal components. Sometimes, our data demands non-orthogonal principal components to represent the data.



The green color vectors are principal components. But, the actual maximum variance directions are blue color vectors. PCA fails to find that vectors. But, Independent Component Analysis (ICA) works well for the above data and it gives the blue color vectors as independent components

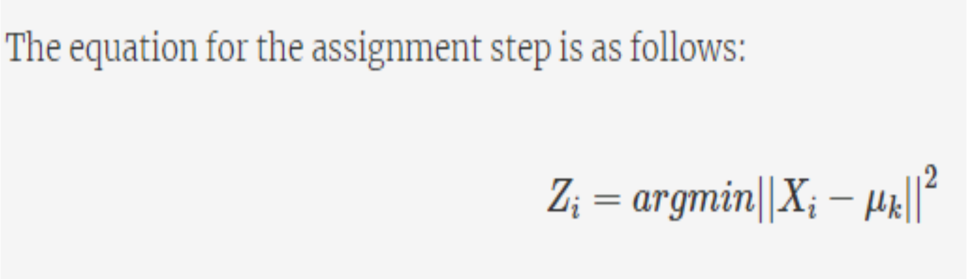
3. PCA assumes that columns with low variance are not useful, which might not be true in prediction setups (especially classification problem with class imbalance)

4. Mean and covariance don't describe some distributions.

   There are many statistics distributions in which mean and covariance doesn't give relevant information of them. In fact, mean and covariance are used (or could be considered important) for Gaussians.

Question 3 Compare and contrast K-means Clustering and Hierarchical Clustering.
Ans -3

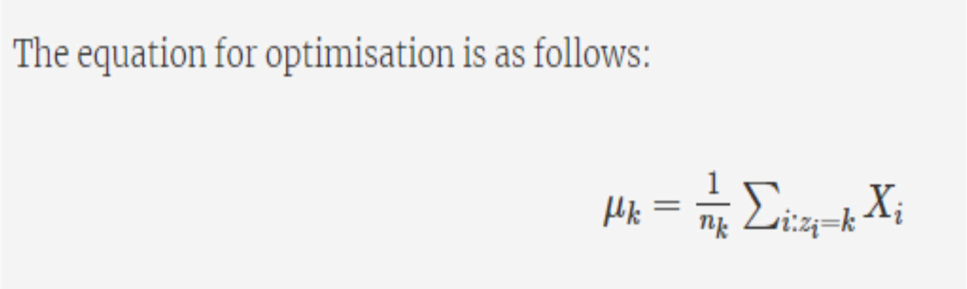
1. Hierarchical clustering is computational intensive so, it cannot be used for large data sets and hence K-Means clustering is recommended for large data. Moreover for large data sets analysing a dendrogram is a pain.
2. Hierarchical clustering has fewer assumptions about data distribution. K Means requires proper separation or cluster tendency, identification of initial centroids and also the suitable value of K before clustering. Hierarchical clustering does not have any such compulsions.
3. The Algorithm of K-Means starts by randomly defining k centroids. From there, it works in iterative (repetitive) steps to perform two tasks:
   - Assign each data point to the closest corresponding centroid, using the standard Euclidean distance. In layman's terms: the straight-line distance between the data point and the centroid.



The equation for the assignment step is as follows:

$$Z_i = argmin||X_i - \mu_k||^2$$

   - For each centroid, calculate the mean of the values of all the points belonging to it. The mean value becomes the new value of the centroid.



The equation for optimisation is as follows:

$$\mu_k = \frac{1}{n_k} \sum_{i:z_i=k} X_i$$

Once step 2 is complete, all of the centroids have new values that correspond to the means of all of their corresponding points. These new points are put through steps one and two producing yet another set of centroid values. This process is repeated over and over until there is no change in the centroid values, meaning that they have been accurately grouped. Or, the process can be stopped when a previously determined maximum number of steps has been met

The algorithm of Hierarchical clustering (Agglomerative)

- At first the no of initial clusters is defined equal to no of observations that is each point belongs to its own cluster
- Then we find the Euclidean distance of each point with each other point.
- Then we fuse the two nearest cluster points to form a new cluster on the basis of their Euclidean distance.
- We continue repeating this step till we fuse together all cluster points to form a single cluster
- This combining of clusters together can be represented by what we call a Dendrogram
- In order to find distance between a combined cluster and single point cluster we look at the Euclidean distance between the single point with all the combined cluster points and choose the minimum distance as the cluster distance.

4. K Means clustering Works well only for round shaped, and of roughly equal sizes/density clusters. Hierarchical Clustering can be used on a wide variety of cluster sizes.