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)

Answer: The dataset has lot of different countries across the different geographical region, the clustering of countries based on their economic, health and other feature which determine their aid requirement
The solution is based out on clustering algorithm using K mean on dimensional reduced data using PCA module.
Step to solution

1) Once EDA is done its seen that there are highly correlated variable which make difficult make insight.
2) Using PCA method, the features are decomposed to PC which covers maximum variance of data/
3) Choosing principle component is based on scree plot which give 5 principle components as it covers 95% of variation in data.
4) Clustering was done on PCA data using both K-mean and hierarchical clustering method , with the help of various visualization technique like elbow curve, silhouette graph and dendrograph cluster has been chosen for further analysis.

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/standardization before performing Clustering.
e) Explain the different linkages used in Hierarchical Clustering.

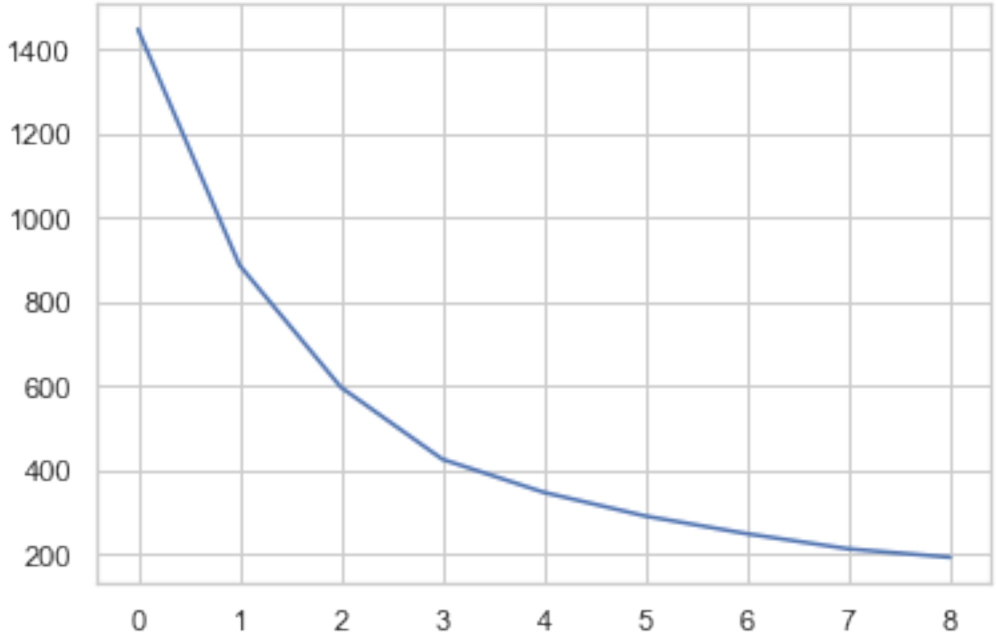
Answer 2:

a) In Hierarchical clustering, clusters have a tree like structure or a parent child relationship. Here, the two most similar clusters are combined together and continue to combine until all objects are in the same cluster.
K- means is a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters. It is a division of objects into clusters such that each object is in exactly one cluster, not several.

b) K-mean clustering is an unsupervised learning technique which cluster data in similar clustering characteristics, below are the steps
   1) Partition the items into K initial clusters, where K is any initial estimate of the number of clusters which can be determined according to the business requirements. Alternatively, it can be determined by using the elbow method (which is a widely used technique)
   2) Euclidean distance with either standardized or unstandardized observations is calculated. Assign an item to the cluster whose centroid (mean) is nearest. Recalculate the centroid for the cluster receiving the new item and for the cluster losing the item.
   3) Repeat Step 2 until no more reassignments take place

c) To find the best k we need to measure the quality of the clusters. The most traditional and straightforward method is to start with a random k, create centroids, and run the algorithm as we explained above. A sum is given based on the distances between each point and its closest centroid. As an increase in clusters correlates with smaller groupings and distances, this sum will always decrease when k increases; as an extreme example, if we choose a k value that is equal to the number of data points that we have, the sum will be zero.

The goal with this process is to find the point at which increasing k will cause a very small decrease in the error sum, while decreasing k will sharply increase the error sum. This sweet spot is called the "elbow point, in the image below, it is clear that the "elbow" point is at *k* = 4 or 5



In business aspect it make sense as after 4 its falling flat which mean that the data values are at same level.

.

d) *Standardization is an important step of Data pre-processing* it controls the variability of the dataset, it convert data into specific range using a linear transformation which generate good quality clusters and improve the accuracy of clustering algorithms, check out the link below to view its effects on k-means analysis.

e) Hierarchical Clustering have 3 different linking methodology

I. Single-Link, : In single-link (or single linkage) hierarchical clustering, we merge in each step the two clusters whose two closest members have the smallest distance (or: the two clusters with the smallest **minimum** pairwise distance).

II. Complete-Link : In complete-link (or complete linkage) hierarchical clustering, we merge in each step the two clusters whose merger has the smallest diameter (or: the two clusters with the smallest **maximum** pairwise distance).

III. Average-Link Clustering : is a compromise between the sensitivity of complete-link clustering to outliers and the tendency of single-link clustering to form long chains that do not correspond to the intuitive notion of clusters as compact, spherical object

IV.

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.

Answer 3:

a) 3 Application of using PCS

i) Network traffic data collected for intrusion analysis is typically high-dimensional making it difficult to both analyze and visualize. Principal Component Analysis is used to reduce the dimensionality of the feature vectors extracted from the data to enable simpler analysis and visualization of the traffic.

ii) Ecological health assessment of a body of water

iii) In quantitative finance, principal component analysis can be directly applied to the risk management of interest rate derivative portfolios

iv) A variant of principal components analysis is used in neuroscience to identify the specific properties of a stimulus that increase a neuron's probability of generating an action potential

b) Important building blocks of PCA:

I. **Basis Transformation** is essentially the same as "**conversion of units"** exercise . Basically, what you do in basis transformation is that you change the representation of the same point from one **"unit" to another which mean you transform in to** new set of basis vectors to represent all the points which we have in our dataset. These basis vectors that we find **explain the information of the dataset in the "best possible way"** and therefore allow us to do operations like dimensionality reduction, find latent variables etc.

II. **Basis Variance** means to extend the maximum variance i.e. The more variance a column has, the more informative it is and the more important it is for our modelling process. Therefore, the ones which explain low variance can be eliminated from our dataset without affecting our results much. This is essentially what dimensionality reduction does .In some cases, the variance might have been equally distributed amongst all the columns What we can do here, in case our standard basis vectors don't explain the dataset in a way which we want i.e. where some of the columns explain far lesser variance than others, thereby making it easier for us to remove them, is to find a newer set of basis vectors which does exactly that. And that gives us the fundamental function of PCA
PCA helps in finding the best possible set of basis vectors for a given dataset in such a way that the variation is non-uniformly distributed amongst them - some columns now explain far more variance than other columns. This makes it easier to choose which columns to keep and which to discard

c) PCA has 3 major assumptions/simplifications embedded –

1. The PCs have to be linear combinations of the original columns • Why limit ourselves to linearity when we can go non-linear? • t-SNE is an alternative, although computationally very expensive

2. PCA requires the PCs to be uncorrelated/orthogonal/perpendicular • Sometimes the data demands that correlated components to represent the data • ICA (Independent Component Analysis) overcomes this drawback, but is several times slower than PCA

3. PCA assumes low variance components are not very useful - • In supervised learning situations, this can lead to loss of valuable information. This is especially true for highly imbalanced classes/variables.