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)

PROBLEM STATEMENT:

- HELP International, an NGO is has just raised 10 Mn USD and is tasked with allocating for development of underdeveloped countries.
- Since they will have to choose between countries to allocate funds appropriately, they were provided with data of 167 countries with various attributes as variables such as GDP per capita, Child Mortality, Income, Inflation etc.
- Our task is to identify a cluster of countries which need funding based on the variables from the datasets.
- Countries which require funding are usually the ones which are doing poorly on the economic indicators. They are usually associated with low GDP per capita, High Child Mortality, Low income, Low inflation, Lower life expectancy etc.
- Based on the above analysis, we need to find a cluster of countries which are doing poorly on the economic front.

METHODOLOGY:

DATA PREPARATION:

- Since there are no missing values in the data. We can proceed directly for Outlier treatment.
- It is observed that all the 8 variables have outliers and they are treated by considering the values above 99%'le as values at the 99th %'le and the values below 1%'le as values at the 1st %'le.
- After the Outliers are treated, all the values are standardized using standard scaler to perform PCA.

RESULTS OF PCA:

- After performing PCA, from the explained variance ratio, it is observed that more than 95% of the variance in the dataset is contributed by 5 principal components and hence dataset is transformed with 5 principal components to perform Kmeans clustering.

Kmeans Clustering:

- Number of clusters in Kmeans clustering bis selected by plotting an elbow curve using sum of squared differences.
- From the curve, it is observed that SSD is drastically reducing as we move from 4 to 5 clusters indicating that the optimal number of clusters is 4.
- Also, Silhoutte number is calculated and it shows a good number above 0 for k = 4 with the silhouette number decreasing as we increase the number of cluster
- Therefore, Kmeans clustering is performed on the dataset with 5 principal components to obtain the clusters.

Results of Kmeans Clustering:

- Clusters thus performed with Kmeans clustering are analyzed with 3 of the original variables i.e GDP per capita, Child Mortality & income.
- From the results of Kmeans clustering, It is clear that countries in the cluster 0 exhibit the characteristics of low income, low GDP per capita and high child mortality.
- Therefore, as per the Kmeans clustering it is advisable to allocate funds for the development of countries in cluster 0.

HEIRARCHIAL CLUSTERING:

- For Heirarchial clustering, a Dendrogram is created using the complete linkage method. A Dendrogram is also created using a single linkage method but due to the limitations of single linkage method, clusters formed are not very clear.
- Dendrogram is cut in such a way that the number of clusters is 4.

RESULTS OF HEIRARCHIAL CLUSTERING:

- After performing Heirarchial clustering, clusters thus formed have been analyzed with respect to principal components and the three chosen original variables.
- Therefore, as per Heirarchial clustering countries from cluster 1 are to be funded for development.

Choosing the countries:

- Finally, as per the analysis, 47 countries require funding as per Kmeans clustering and 10 countries require funding as per Hierarchical clustering.
- It is also observed that 8 countries in the Hierarchical clustering also are present in the list of countries from Kmeans countries incdicating that these countries are in the direst need of funds.
- Therefore, the final list of countries are common in the Kmeans clustering and Hierarchical clustering. Final 8 countries are as follows.

1. Angola
2. Congo Rep
3. Equitorial Guinea
4. Mauritania
5. Nigeria
6. Sudan
7. Timor - Leste
8. Yemen

QUESTION 2:

a) Compare and contrast K-means Clustering and Hierarchical Clustering.

KMeans	Hierarchical
Number of clusters is chosen before forming the cluster	Number of clusters is chosen after forming the clusters
Data is divided into clusters in the 1st step itself and later iterated to optimise the clusters by identifying centroids.	Data is merged with other data points in a series of steps starting from forming a cluster of of 2 data points and finally forming a single cluster containing all the data points

Since clusters are formed in the 1st step and later iterated to optimise, computing power required is low	Since distance from each point to every other point needs to be calculated, computing power required is high

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

- In the Kmeans clustering, the number of clusters is decided initially. Let the number of clusters be n.
- In the plane containing data points , n random points are chosen initially.
- Each data point is initially assigned to one of the n cluster based on the nearest cluster to the data point calculated by Euclidean distance.
- Once, the initial clusters are performed, a centroid which is the mean of coordinates of the data points is formed for each cluster.
- Now, the clusters are again formed using the minimum distance between centroids and data points.
- The above process is iterated multiple times until the clusters don't change anymore on further iterations.

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

STATISTICAL ASPECT:

An optimal cluster is where the distance between points within the cluster is minimum and distance between the points in 2 different clusters is maximum.

In order to ensure this, there are essentially 2 methods that are followed. They are:

1. Elbow cuve
2. Silhoutte method

Elbow curve:

Here a curve is formed by considering different number of clusters and plotting a sum of squared distances between the points in different cluster against number of clusters. By increasing the number of clusters, obviously the value of SSD decreases but the rate of that will be drastically once the number of clusters is increased beyond optimal number of clusters.

Silhouette method:

To compute silhouette metric, we need to compute two measures i.e. p and q where,
p is the average distance from own cluster(Cohesion).
q is the average distance from the nearest neighbour cluster(Separation).

Silhouette number is calculated by formula = p-q/max(p,q)

This shows that a model containing silhouette number approaching 1 is optimal.

BUSINESS ASPECT:

In some cases, the silhouette number may be high for a very low number of clusters but it does not make a business sense to form lesser clusters. For example, a client may want to segment newly acquired customers into 4 different clusters to analyse the data. In that case, number of clusters may be considered as 4 inspite of a lower silhouette number.

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

In the clustering process, each data point consists of multiple variables of varying magnitudes and clustering is carried out by calculating distance between data points. Therefore, when the scaling/standardization is not done, it is possible that we may see large fluctuations in these distances because of overweighting one variable in the dataset. This causes huge disruption in the formation of clusters and hence, to avoid this, scaling is performed.

e) Explain the different linkages used in Hierarchical Clustering.

- Single Linkage: Here, the distance between 2 clusters is defined as the shortest distance between points in the two clusters
- Complete Linkage: Here, the distance between 2 clusters is defined as the maximum distance between any 2 points in the clusters
- Average Linkage: Here, the distance between 2 clusters is defined as the average distance between every point of one cluster to every other point of the other cluster.

QUESTION 3:

a) Give at least three applications of using PCA.

PCA is basically dimensionality reduction technique which is used to eliminate a lot of features in dataset that do not contribute to the variance in the dataset. Following are the main applications.

- For data visualisation and EDA
- For creating uncorrelated features that can be input to a prediction model
- Finding latent themes in the data
- Noise reduction

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

BASIS TRANSFORMATION:

- Basis is essentially the fundamental units in which you express your data.

- In vectors and vector spaces, we use basis vectors to represent the points in space.
- Using the analogy of basis as a unit of representation, different basis vectors can be used to represent the same observations using the following equation

New basis = M * old Basis

VARIANCE:

- Variance is the variation of each variable in the dataset.
- A variable which has more variance in the dataset contributes more to the variation in the dataset and hence contains more information.
- Therefore, dimensionality reduction can be achieved by removing those variables which have negligible variance.

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

- PCA is limited to linearity which means it cannot be applied on non linear models.
- PCA needs the components to be perpendicular, which may not be optimal in some cases. The alternative technique is to use Independent Components Analysis.
- PCA assumes that columns with low variance are not useful, which might not be true in prediction setups such as predicting a fraudulent credit card transaction etc.