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:

Problem Statement: HELP International, an international humanitarian NGO raised the funds and want to invest it for overall development of the countries that are in the direst need of aid. we need to categories the countries using some socio-economic and health factors that determine the overall development of the country.

solution methodology:

1) Perform EDA on the dataset. Read the data, find missing values and do the Outlier Analysis:

2) Perform PCA on the dataset and obtain the new dataset with the Principal Components. Choose the appropriate number of PC components based on the result from scree plot. As we know that 95% variance is explained by first few PC components. We can get that by explained_variance_ratio_. In this case no. of PC s was 5.

3) perform the clustering activity on this new dataset, i.e. the PCA modified dataset with the k components.

4) Try both K-means and Hierarchical clustering (both single and complete linkage) on this dataset to create the clusters.

5) Analyse the clusters and identify the ones which are in dire need of aid by comparing how these three variables - [gdpp, child_mort and income] vary for each cluster of countries to

recognise and differentiate the clusters of developed countries from the clusters of under-developed countries.

6) perform visualisations on the clusters that have been formed.

7) The final list of countries depends on the number of components that are choosen and the number of clusters that are finally formed.

Summary: In this case we got 5 PC components and Kmean clustering was performed with 3 clusters(Based on the Elbow curve). These 3 clusters were Developed, underdeveloped and developing countries.

Both Kmean and Hierarchical clustering methods are used but I found Kmean method more efficient as we are sure how many clusters need to be formed based on elbow curve graph but in Hierarchical clustering. It was not sure as to where to cut the tree but I went ahead with 3.

I took the mean of GDPP, Child_mort and Income for the cluster 0(underdeveloped) and then took out the list of countries which were less then average on GDPP, Child_mort and Income

From the original database and then selected 5 countries in the increasing order of GDPP. As I understand GDPP of a country reflect the growth of country and is an important factor to consider.

Question 2: Clustering

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

Ans

With k-Means clustering, you need to have a sense ahead-of-time what your desired number of clusters is (this is the 'k' value). Also, k-means will often give unintuitive results if (a) your data is not

well-separated into sphere-like clusters, (b) you pick a 'k' not well-suited to the shape of your data, i.e. you pick a value too high or too low, or (c) you have weird initial values for your cluster centroids (one strategy is to run a bunch of k-means algorithms with random starting centroids and take some common clustering result as the final result).

In contrast, hierarchical clustering has fewer assumptions about the distribution of your data - the only requirement (which k-means also shares) is that a distance can be calculated each pair of data points. Hierarchical clustering typically 'joins' nearby points into a cluster, and then successively adds nearby points to the nearest group. You end up with a 'dendrogram', or a sort of connectivity plot. You can use that plot to decide after the fact of how many clusters your data has, by cutting the dendrogram at different heights. Of course, if you need to pre-decide how many clusters you want (based on some sort of business need) you can do that too. Hierarchical clustering can be more computationally expensive but usually produces more intuitive results.

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

Ans

K-Means algorithm is the process of dividing the N data points into K groups or clusters. Here the steps of the algorithm are:

1. Start by choosing K random points the initial cluster centres.
2. Assign each data point to their nearest cluster centre. The most common way of measuring the distance

   between the points is the Euclidean distance.

3. For each cluster, compute the new cluster centre which will be the mean of all cluster members.

4. Now re-assign all the data points to the diffrent clusters by taking into account the new cluster centres.
5. Keep iterating through the step 3 & 4 until there are no further changes possible.

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

1. Elbow method:-

- Compute clustering algorithm (e.g., k-means clustering) for different values of k. For instance,

  by varying k from 1 to 10 clusters.

- For each k, calculate the total within-cluster sum of square (wss).
- Plot the curve of wss according to the number of clusters k.
- The location of a bend (knee) in the plot is generally considered as an indicator of the appropriate number of clusters.

2. Average silhouette Method

- Compute clustering algorithm (e.g., k-means clustering) for different values of k. For instance,

  by varying k from 1 to 10 clusters.

- For each k, calculate the average silhouette of observations (avg.sil).
- Plot the curve of avg.sil according to the number of clusters k.

- The location of the maximum is considered as the appropriate number of clusters.

business :
From the business perspective we need to see how many relevant categories data should be divided into. We need that many no. of clusters

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

Standardisation of data, that is, converting them into z-scores with mean 0 and standard deviation 1,

is important for 2 reasons in K-Means algorithm:

- Since we need to compute the Euclidean distance between the data points, it is important to ensure that the attributes with a larger range of values do not out-weight the attributes with smaller range. Thus, scaling down of all attributes to the same normal scale helps in this process.
- The different attributes will have the measures in different units. Thus, standardisation helps in making the attributes unit-free and uniform.

e) Explain the different linkages used in Hierarchical Clustering.
Ans:

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: Principal Component Analysis

a) Give at least three applications of using PCA.

Ans

PCA is predominantly used as a dimensionality reduction technique in domains like facial recognition, computer vision and image compression. It is also used for finding patterns in data of high dimension in the field of finance, data mining, bioinformatics, psychology, etc.

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

Ans:

1) **Basis Transformation** is essentially the same as **"conversion of units"** exercise which we do in your primary school. For example, conversion of rupees to dollars and vice-versa, conversion of inches to cm and vice-versa and so on. Basically, what we do in basis transformation is that we change the representation of the same point from

one **"unit" to another.** Like in currency, the amount 1500 rupees can be represented as 22 US Dollars or 19 Euros or 2300 Yen and so on.

In our context, the fundamental thing that PCA helps us to do is find a 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.

2) variance as information

the variation present in a column is an indicator of how important that column is for our modelling setup. Hence we use the following rule: if the column has high variance, then that column is important for our modelling process. Else, if the variance is quite low as compared to others, then we can disregard that column and use other columns for our modelling purposes. This in its essence is how we are going to do **dimensionality reduction in PCA.**

Combining the 2 Building Blocks

1) Basis Transformation allows us to represent the same data in multiple basis vectors.
2) 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.

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

Ans

1) PCA is limited to linearity, though we can use **non-linear techniques such as t-SNE** as well (you can read more about t-SNE in the optional reading material below)
2) PCA needs the components to be perpendicular, though in some cases, that may not be the best solution. The alternative technique is to use **Independent Components Analysis**
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)