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 #1

Problem Statement

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 programmes, 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.

Solution Approach

The steps are broadly:

1. *Read and understand the dataset provided*
2. *Clean the data*
3. *Outliers analysis and treatment*
4. *Scale the features*
5. *Perform PCA(Principal Component Analysis) on the data*
6. *Clustering*
    i. *K-Means Clustering*
    ii. *Hierarchical Clustering*
7. *Analyse and Visualise the principal components*
8. *Analyse and Visualise the original variables*
9. *Final Analysis and recommendations*

After performing PCA, the first 4 variables, explained the ***95%*** *of variance, hence, 4 variables also known as Principal Components were taken for the further analysis.*

Hierarchical clustering produces better results. After performing complete linkage, dendrogram cut the edges across clusters or groups, hence, divided the countries into 5 clusters.

On further analysis of GDPP, Income and Child Mortality rate, the conclusion is that the countries belong to cluster ‘0’ have low GDPP, low income and very high child mortality rate. Hence, the NGO must focus more on these countries. List as follows:

Afghanistan
Benin
Botswana
Burkina Faso
Burundi
Cameroon
Central African Republic
Chad
Comoros
Congo, Dem. Rep.
Cote d'Ivoire
Eritrea
Gambia
Guinea-Bissau
Haiti
Iraq
Kenya
Kiribati
Lao
Lesotho
Liberia
Madagascar
Malawi
Mali
Micronesia, Fed. Sts.
Mozambique
Namibia
Niger
Pakistan
Rwanda
Senegal
Sierra Leone
Solomon Islands
South Africa
Tanzania
Timor-Leste
Togo
Uganda
Zambia

Question 2*:* Clustering

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

Answer 2a:

1. *In K-means clustering, we have to choose the initial cluster centres, whereas Hierarchical clustering builds clusters within clusters and does not require a pre-specified number of clusters*
2. *In k-means clustering, we try to identify the best way to divide the data into k sets simultaneously. A good approach is to take k items from the data set as initial cluster representatives, assign all items to the cluster whose representative is closest, and then calculate the cluster mean as a new representative; until it converges. While Hierarchical clustering has two approaches to divide the data into clusters, they are top-down and bottom-up.*

   In top-down hierarchical clustering, we divide the data into 2 clusters (using k-means with k=2k=2, for example). Then, for each cluster, we can repeat this process, until all the clusters are too small or too similar for further clustering to make sense, or until we reach a preset number of clusters.

   In bottom-up hierarchical clustering, we start with each data item having its own cluster. We then look for the two items that are most similar, and combine them in a larger cluster. We keep repeating until all the clusters we have left are too dissimilar to be gathered together, or until we reach a preset number of clusters.

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

Answer 2b:

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 different 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.

Answer 2c:

There are a number of pointers that can help us decide the K for our K-means algorithm:-

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.*

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

Answer 2d:

Standardisation of data, that is, converting them into z-scores with mean 0 and standard deviation 1,is important for 2 reasons in clustering:

1. *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.*

2. *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.

Answer 2e:

There are three types of linkage available 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*:* Principal Component Analysis

a) Give at least three applications of using PCA.

Answer 3a*:*

PCA is predominantly used as a dimensionality reduction technique in domains like facial recognition, computer vision and image compression.

1. ***The predictive model***

Having a lot of correlated features lead to the multicollinearity problem. Iteratively removing features is time-consuming and also leads to some information loss.

2. ***Data visualisation:***

It is not possible to visualise more than two variables at the same time using any 2-D plot. Therefore, finding relationships between the observations in a data set having several variables through visualisation is quite difficult.

3. ***Data Preprocessing:***

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

Answer 3b*:*

Basis Transformation:

Essentially, 'basis' is a unit in which we express the vectors of a matrix.

For example, we describe the weight of an object in terms of kilogram, gram and so on; to describe length, we use a metre, centimetre, etc. So for example, when you say that an object has a length of 23 cm, what you are essentially saying is that the object's length is 23×1 cm. Here, 1 cm is the unit in which you are expressing the length of the object.

Similarly, vectors in any dimensional space or matrix can be represented as a linear combination of basis vectors.

Variance as information :

The variance as information measures the importance of a column by checking its variance values. If a column has more variance, then this column will contain more information.

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

Answer 3C:

Shortcomings of PCA:

1. *PCA is limited to linearity, though we can use non-linear techniques such as t-SNE as well .*
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 a high class imbalance).*