Assignment Part:2

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 :** An international humanitarian NGO (HELP International) 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. Now the CEO of the NGO needs to decide how to use $ 10 million money strategically and effectively. So as to better utilize the funds an analysis was required to categorise the countries using some socio-economic and health factors that determine the overall development of the country.
- **Solution Methodology** : A total of 167 countries were under analysis. And all these countries have 9 features covering the crucial aspects of information about the countries. Steps that were followed are:
  - **Data Cleaning** : The data was already in good condition and it had no missing values or duplicates
  - **Applying PCA** : Removed multicollinearity from the data so as to obtain independent variables. Overall we choose 6 components because they were covering more 95% of information. Without loosing much information we dropped 3 components. Before applying PCA we also standardised the data.

- **Dropping Outliers** : Since we had to provide an aid to countries which are extremely poor. We can remove statistical 5% of countries which do not have problems of any kind.
- **Creating clusters** : Used Hierarchical clustering (Single and complete) and K-Means clustering. Tried to work out with number of clusters equal to 3 and 5. Clusters formed with K-Means were not able to differentiate amongst each other. Hierarchical with single linkage was also not able to create better clusters. The clusters formed kmeans had similar means and similar spread of data in every cluster.

  Finally Hierarchical clustering(complete) created clusters which were visually understandable.
- We also plotted the countries in world map. which gave us an outlook that of how countries look in map with respect to labels. And African countries were mostly that may require an aid in case of disaster.

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/standardisation before performing Clustering.

e) Explain the different linkages used in Hierarchical Clustering.

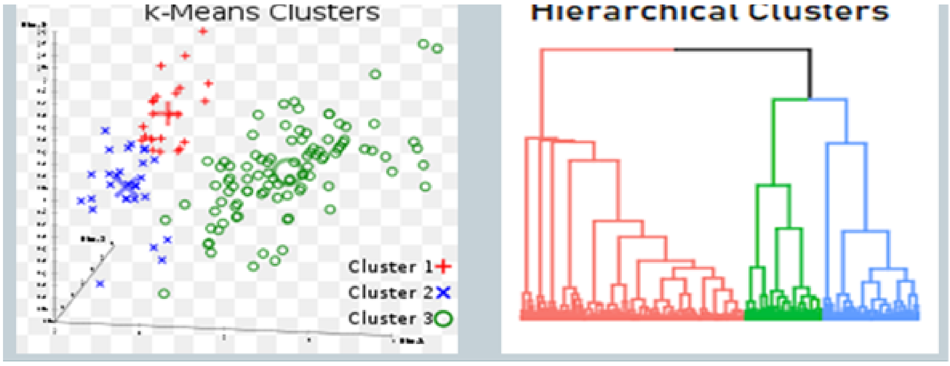
Answer 2)

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

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.

The Algorithm is Resource intensive. Require more memory and computing power.

More easy to identify optimal cluster compared to K-Means algorithm.



K means is an iterative clustering algorithm that aims to find local maxima in each iteration. The goal is to reach global optima. When there will be no further switching of data points between two clusters for two successive repeats. It will mark the termination of the algorithm if not explicitly mentioned.

It is less Resource intensive compared to Hierarchical cluster.

Optimal number of clusters might be inefficient from business prespective.

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

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

Here the steps of the algorithm are:

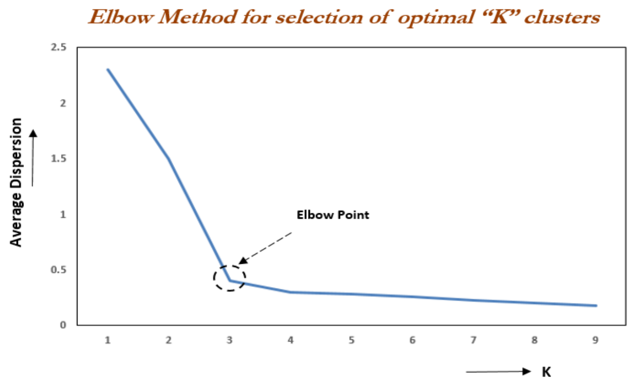
1. Start by choosing K random points the initial cluster centres.
2. Assign each data point to their nearest cluster centre. The most common way of measuring the distance between the points is the Euclidean distance.
3. For each cluster, compute the new cluster centre which will be the mean of all cluster 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.

At this point, you arrive at the optimal clusters.

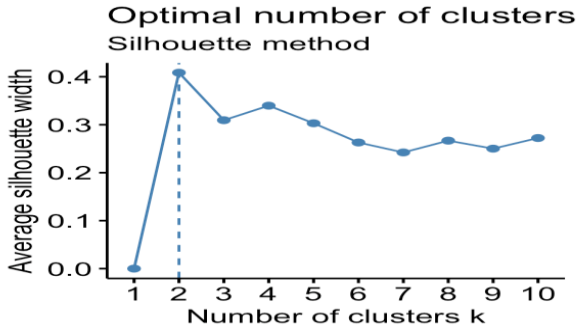
c) **How is the value of ‘k’ chosen in K-means clustering? Explain both the statistical as well as the business aspect of it.**

There are a number of ways 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



When business perspective is considered then number of clusters should be such that they bring some value in terms of enhancement or to the customer.

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

Standardisation of data 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

There are three types of linkages:-

- 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.

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)

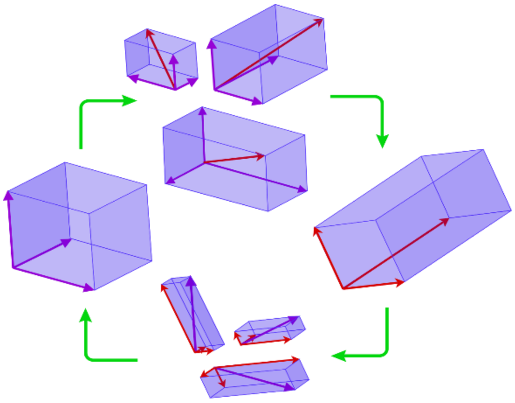
a) Give at least three applications of using PCA.

- **Image Compression** : When we reduce the number of features in an image, the number of components decreases. And the image is still able to preserve the important features.
- **Extremely large number of dimensions**: When number of features are so much that it is physically impossible to pick important features.
- **Neural Networks** : When number of features are reduced then then models converge fast as compared to otherwise.

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

Basis transformation is the process of converting your information from one set of basis to another. Or, representing your data in new columns different from the original. Often for convenience, efficiency, or just from common sense.

Data does not change but viewing angle is changed.



variance as information :Variables which capture variance in the data, are the variables that capture the information in the data! Taking the idea further, if two variables are very highly correlated, they together don’t add a lot information than they do individually

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

PCA has 3 major assumptions/simplifications embedded –

- The PCs have to be linear combinations of the original columns. Principal Components are not as readable and interpretable as original features.
- PCA requires the PCs to be uncorrelated/orthogonal/perpendicular • Sometimes the data demands that correlated components to represent the data
- PCA assumes low variance components are not very useful
  - Although Principal Components try to cover maximum variance among the features in a dataset, if we don't select the number of Principal Components with care, it may miss some information as compared to the original list of features
- Data standardization is must before PCA: You must standardize your data before implementing PCA, otherwise PCA will not be able to find the optimal Principal Components.