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:

- Reading and understanding of dataset ‘country-data.csv’
- Checking for missing values and outliers
- Treating Outliers (To get unbiased result)
- Standardization of continuous variables are done to enhance the computation power
- Checking correlation
- PCA is performed as correlation existed among variables and by PCA we can reduce variables without losing too much data (less variables leads to better performance and less time for analysis)
- 5- number of components are taken as from the scree plot, it is found that these 5 number of components can describe 95% variance in the data set
- Correlation is checked again and ‘0’ correlation is found
- K-means clustering is performed with k=3 (randomly taken)
- From the SSD and silhouette analysis analysis k is taken as 4
- Hierarchical Clustering is also performed with k=4
- Results of both k-means and Hierarchical Clustering are compared and it is found that k-means clustering gives better results (analysis is done by using boxplots)
- Final top-10 countries are suggested

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:

a) Comparison between K-means and Hierarchical clustering

- K-means clustering aims to find local maxima in each iteration while Hierarchical clustering builds hierarchy of clusters
- Hierarchical clustering can not handle big sized data properly while K-means clustering can

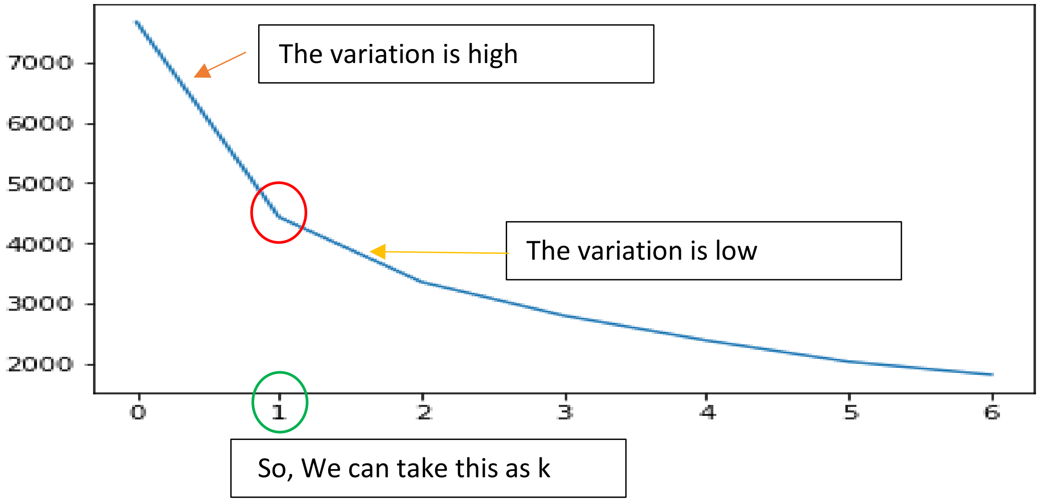
- In K-means clustering, we provide initial number of clusters before iteration while in Hierarchical clustering, all the data points are assigned to a cluster of their own.

b) Steps of K-means clustering

- Start by choosing k random cluster centers
- Assign each point to their nearest cluster centre. Euclidean distance method is used to measure the distance
- For each cluster, compute the new cluster centre which will be the mean of all cluster members.
- Now reassign all data points to these new cluster centres according to their distance from the nearest cluster centre
- Keep iterating until there are no further changes possible

c) The value of k in K-means clustering can be selected by using either SSD/Elbow curve or silhouette analysis or both

SSD Analysis



The variation is high

The variation is low

So, We can take this as k

Silhouette Analysis

Remarks:

The value of the silhouette score range lies between -1 to 1.

A score closer to 1 indicates that the data point is very similar to other data points in the cluster,

A score closer to -1 indicates that the data point is not similar to the data points in its cluster.

Example:

For n_clusters=2, the silhouette score is 0.6415858
For n_clusters=3, the silhouette score is 0.6084896
For n_clusters=4, the silhouette score is 0.5814786
For n_clusters=5, the silhouette score is 0.5658529
For n_clusters=6, the silhouette score is 0.5170796
For n_clusters=7, the silhouette score is 0.5158077
For n_clusters=8, the silhouette score is 0.5059904

So, We can take this as k

d) **Necessity for scaling/standardization before performing Clustering:**

- Scaling or standardization rescales the values of the variables into a common scale
- Scaling is effective when dataset contains variables of different units (km, inches, number, etc.) because groups are defined based on the distance between points.
- Scaling also enhance the computation power as all values of variables share a common scale (i.e. calculation between 5 and 10 is faster than 5 and 4958).

e) **Types of linkage are :**

I) **Single Linkage:**

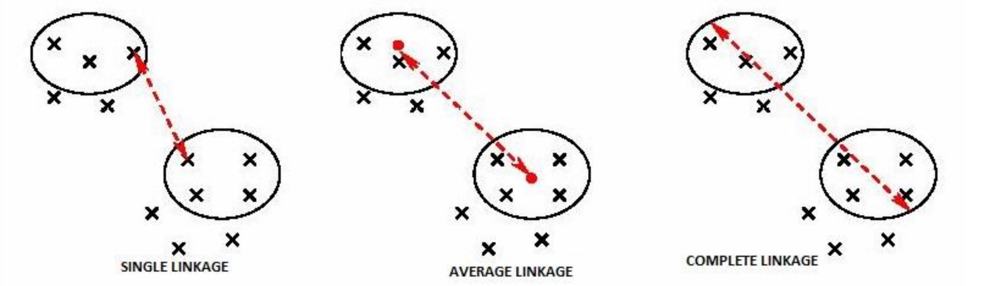
The distance between 2 clusters is defined as the shortest distance between points in the two clusters

II) **Complete Linkage:**

The distance between 2 clusters is defined as the maximum distance between any 2 points in the clusters

III) **Average Linkage:**

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:

a) **Applications of using PCA:**

1) Image compression
2) Disease control
3) Neuro Science
4) Finance

b) v

c) **Shortcomings of using Principal Component Analysis:**

1) Data standardization is must before PCA
2) Independent variables become less interpretable
3) Loss of data when number of PCA component is selected carelessly