Question 1: Assignment Summary

Answer:

Problem Statement: We need to categorize the countries using some social-economic and health factors that determine the overall development of the country. Then we need to find out the final list of the countries which require aid?

Solution: We first scale the three columns i.e. imports, exports and health to absolute values according to their GDP per capita. Then we rescale the variables using standard scalar so that PCA can be applied, next step is the plotting of scree plot, through which we found that over 90% of the data is properly explained by the first three principal components thus we created the matrix according to the given three principal components. Next step is to remove the outliers before clustering. In clustering we calculate the Hopkins statistics to ensure that the data is good for clustering, we got the Hopkins statistics as 0.778. Based on that we did two types of clustering K-means and Hierarchical clustering. By silhouette score in K-Mean clustering we found that there are 5 number of clusters in which 2 and 4 are most appropriate. Then in Hierarchical clustering we first used linkage with single and complete method, and dendrogram to represent clear clusters, by which we found that clusters 0 and 1 are more appropriate.

Next part is to find out the final list of countries which require aid. This is done on the basis of three factors used in part-1 too i.e. child mortality, income and gdpp. According to the analysis in first part the child mortality should be at least 76, income should be at least 3200 and gdpp should be less than 1700. After analysis, it is found that there are 23 countries which require aid.

Question 2:

a) **Compare and contrast K-means Clustering and Hierarchical Clustering.**
   Answer:
   K Mean clustering is the clustering in which the data is partitioned into subsets. Each cluster is a subset in which the similarity between a cluster is more and similarity between the clusters is less.
   In Hierarchical clustering hierarchical decomposition of the given set of data is created. It is created two-way, from bottom to top and from top to bottom. These two approaches are known as agglomerative approach and divisive approach.
   In K Mean clustering the clusters are defined first while in hierarchical clustering the clusters are derived after analysis.
b) **Briefly explain the steps of the K-means clustering algorithm.**
   Answer:
   Step 1: Specify the number of clusters.
   Step 2: Then assign each data point to a cluster randomly.
   Step 3: Next we need to compute the cluster centroid
   Step 4: Then reassigning each point to the closest cluster centroid.
   Step 5: Re computing the cluster centroids.
   Step 6: Repeat step 4 and 5 till we reach same results.
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:

K clusters are specified randomly. There can a direct approach or statistical testing method. The direct method can be the sum of squares or the average silhouette within the cluster. While the statistical testing method consist of comparing evidence against null hypothesis. Elbow method is also used to find the appropriate number of clusters.

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

Answer:

It is important to scale the data and then converting it into standard form because different components have different features so to remove the inconsistency of the data we need to first scale then data then standardise.

e) Explain the different linkages used in Hierarchical Clustering.

Answer:

Different linkages used in clustering are single, complete and average.

In Single linkage clustering we merge the two clusters which has the smallest distance. While in complete linkage clustering we merge two clusters which has the smallest diameter or we can say that merge clusters with maximum pairwise distance. In average linkage clustering the distance between two clusters is defined as the average of distance between the object pairs.

Question 3:

a) Give at least three applications of using PCA.

Answer:

First and the main application of using PCA is for dimensionality reduction for example, image compression, facial recognition and computer vision.

Secondly it can be used to find patterns in data of high dimensions like in banking sector, finance sector or may be in psychology and biometrics.

Thirdly, it can be used to detect and visualize the computer network attacks.

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

Answer:

Basis Transformation: In basis transformation, standardization of the data points is must to transform them into data points expressed in Eigen vectors which will then leads to the dimensionality reduction. The choice of basis transforms our co-variance matrix into diagonal form in which the diagonal element represents the variance on each axis.

Variance as information: In PCA we calculate co-variance matrix from different derived principal components. PCA is designed to maximize the variance of the first k components and to minimize the variance of last p-k components. If the variance is more than the amount of information the variance contain is also more. When we perform dimensionality reduction we correlate the transform predictors and keep their variance to original.

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

Answer:

1) Data Standardization is mandatory before PCA.
2) We can face information Loss during PCA because PCA tries to cover maximum variance among the features in the dataset.

3) After performing PCA our independent variables become less interpretable.