Question 1: Assignment Summary

Ans:-

Problem Statement:-

- A 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. After a funding program the company able to raise $10M now the company want to distribute the money to direst need of aid. We need to analyse the data and suggest the countries by overall which are direst needed aid.

Solution Methodology:-

- The Problem statement signifies that it is a problem which can be solved by segmentation of countries by performing clustering so we need to approach the clustering techniques to achieve the best results.
- First we need to Read the data set and clean the data set look out for null values and convert the % values to numerical values.
- Did Outlier analysis by plotting boxplots for features needed most for the clustering
- Did some basics EDA in features like child_mort , gdpp and income did Univariate and bivariate analysis draw insights .
- To build a model we need to scale the data to a Standard scale so that all the data having different units lies inside a range .
- Used Hopkins method to detect that is the dataset can be able to form clusters or not.
- We need to pre define the cluster for the analysis so we did Elbow curve and Silhouette score analysis and define the no. of cluster required.
- We applied first Kmeans algorithm and forms the cluster and analyse the cluster by plotting and it seems cluster 2 is the most needed aid countries chosed top 10 most needed countries.
- We applied Hierarchical Clustering algorithm applied both in Single and Complete linkage and forms dendrogram .Cut the dendrogram take the value of no. of cluster and Findout that cluster-0 having mostly needed aid .
- By considering both the clustering algo rithm we conclude top 10 countries which are direst needed aid.

Question 2: Clustering

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

Ans:-

K-Means Clustering:-

- This is the process of dividing N data points to K groups of clusters.
- Start by choosing a K random points the initial cluster centres.
- Assign each data point to their nearest cluster centre. The most common way of measuring the distance between the points is the Euclidean distance .
- For each cluster, compute the new cluster centre which will be the mean of all cluster members.
- Now re assign all the data points to the different clusters by taking into account the new cluster centres.
- Keep iterating through step 3 and 4 until there are no further changes possible.

Hierarchical Clustering:-

- Calculate the NxN distance (similarity) matrix, which calculates the distance of each data point from the other.
- Start by assigning each item to its own cluster, so that if you have N items, you now have N clusters, each containing just one item.
- Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now you have one less cluster.
- Compute distances (similarities) between the new cluster and each of the old clusters.
- Repeat steps 3 and 4 until all items are clustered into a single cluster of size N.

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.
- Assign each data point to their nearest cluster centre. The most common way of measuring the distance between the points is the Euclidean distance.
- For each cluster, compute the new cluster centre which will be the mean of all cluster members.
- Now re-assign all the data points to the different clusters by taking into account the new cluster centres.
- Keep iterating through step 3 and 4 until there are no further changes possible.
- Some considerations- The choice of initial cluster centre has an impact on the final cluster composition

- Choosing the number of clusters K in advance- By plotting the datas on Elbow curve and Silhoutte score we can find out the K value which is need to consider before the analysis.
- Outlier can impact on the data clustering.
- Standardisation of data is must needed to convert all large scale datas to a range of std deviation 1 and mean as 0.
- This algorithm is not applicable for the categorical datas.

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

Ans:-

Chosing K on basis of business aspect:-

- According to the business need K should be chose ,As how many no. of segments needed to solve the problem. Lets take a example of a e-commerce site flipkart. Flipkart want to give discounts on the basis of there purchase of a customer frequently so .The problem is team decide to segment the all customer to 3 categories Gold,Silver,Bronze so that the marketing team can release the offer to them individually by there segment.So here we need to make 3 cluster.

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

Statistical method of Selecting K- value:-

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.

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,

Ans:-

- 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
- It controls the variability of the dataset, it convert data into specific range using a linear transformation which generate good quality clusters and improve the accuracy of clustering algorithms.

e) Explain the different linkages used in Hierarchical Clustering.

Ans:-

- Single Linkage:-Distance between 2 clusters is defined as the shortest distance between points in the 2 cluster. Because of the intra cluster linkages it is so hard to interpret the results.
- Complete Linkage:-Distance between two clusters is defined as the maximum distance between any 2 points in clusters. This will form a table and clean form of linkages between the clusters data points.
- Average Linkage:-Distance between 2 clusters is defined as the average distance between every point of one cluster to the every other point of the other cluster.