Question 1: Assignment Summary

Answer:-In the clustering Assignment, I have to find out the top five countries which are in the direst need of the aid. In the starting part of the assignment, I first import warnings and the necessary libraries that are required for the clustering assignment. I import the data and read the data. After importing the data, I did the data Inspection. In the data inspection, I check the shape of the data, data types of the columns and statistical description of data i.e. count, mean, 25^{th} Percentile, 50^{th} percentile, 75^{th} percentile, minimum, maximum, etc. In the data cleaning part, I checked the missing values in the data. The data didn't contain the missing values.

The next step is Exploratory Data Analysis. In the Exploratory Data Analysis, I performed Univariate Analysis, Bivariate Analysis and Multivariate Analysis. In the Univariate Analysis, I visualize data using Boxplots. In the Bivariate Analysis, I visualize the data using Scatterplot, barplot and horizontal barplot. In the Multivariate analysis, I visualize the data using the pairplot and the heatmap. After that in the Outlier Handling step, I perform the capping of the outliers. I cap the outliers less than 1 percentile with 1 percentile value. And cap the outliers greater than 99 percentile with 99 percentile value. I perform the capping of outliers for the 'child_mort', 'exports', 'income', 'inflation' and 'gdpp' columns.

The next step I performed is Hopkins Statistics Test. The Hopkins Statistics test should be greater than 80. For me it is 88. I scale the data. I perform MinMaxScaling on the data. In the MinMaxScaling, the whole data is compressed within 0 and 1.

The model building process starts. For the K-means clustering, the value of k should be determined. There are two ways to calculate the values of k. These are Silhouette score and Elbow curve. After performing these two methods, I choose the value of k is 3. I perform the k-means clustering. Visualize the kmeans clustering and perform cluster profiling on it. After that I find out the top 5 countries which are in the direst need of aid. These countries are 'Sierra Leone', 'Haiti', 'Chad', 'Central African Republic', 'Mali'.

I perform Hierarchical clustering. In the hierarchical clustering, there are two methods, single linkage and complete linkage. In this assignment, I used complete linkage for the hierarchical clustering. And made clustering by taking the no. of clusters as 3. Then, I perform the visualization of hierarchical clustering and performed cluster profiling. After that I find out the top 5 countries that are in the direst need of aid. These are the same countries as in the kmeans clustering.

Question 2: Clustering

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

Answer:-Hierarchical clustering can't handle big data well but K Means clustering can. Because we can divide the data into multiple parts after knowing the no. of clusters i.e. K. But, we can choose any number of clusters you for hierarchical clustering by look at the dendrogram

In K Means clustering, since we can start with random choice of clusters, the results produced by running the algorithm multiple times might show different result. While results are change every time when we change the no. of clusters in Hierarchical clustering.

K Means clustering needs to specify the no. of clusters. But in hierarchical clustering, we don't need to specify the no. of clusters.

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

Answer:- K-means is one of the simplest unsupervised learning algorithms for solving the clustering problems. Followings are the necessary steps for Kmeans clustering algorithm.

1) Randomly select the cluster centers. There are several data points present in the graph. We have to choose any random points as cluster centers.

2) Calculate the distance between each data point and cluster centers.

3) Assign the data point to the cluster center whose distance from the cluster center is minimum. This process need to be done for every cluster centers.

4) Repeat the process again and again. Do it till no more changes are found in the analysis.

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: - The method that performed the process for choose the value of 'k' known as elbow method in KMeans Clustering Algorithm. After plotting the elbow curve, we can

See the value of 'k' is significantly changes for each point. As the value increases in 'k', the graph shows downfall each data point. So we can consider that point as the value of 'k' where the graph shows the higher downfall between the data points.

As we consider the business aspect, the value of 'k' shows that there will be no change in the clustering of group and we can find the similarity between the data points in the higher points. So we have to consider the value of 'k' at that particular given point.

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

Answer:-For the nature of data we have to perform the scaling/standardization before performing Clustering. We can get various types of data for clustering. Some of them are quite large others are either medium range of data or low range of data. So we have to perform outlier capping, for which we can get right amount of data for clustering. In order to get rid of the outliers present in the data, capping is an excellent technique. In the scaling process, we can choose some percentile of the data after the capping so that the results are quite good after clustering.

e) Explain the different linkages used in Hierarchical Clustering.

Answer:- There are 3 types of linkage used in Hierarchical clustering.

1) **Single Linkage:** Here, the distance between 2 clusters is defined as the shortest distance between points in the two clusters
2) **Complete Linkage:** Here, the distance between 2 clusters is defined as the maximum distance between any 2 points in the clusters
3) **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.