SUBJECTIVE QUESTIONS ON CLUSTERING ASSIGNMENT

QUESTION 1: ASSIGNMENT SUMMARY

Problem Statement: The international organisation, HELP NGO was successful in raising a fund of $10 million and the CEO has decided to invest the amount in the backward countries inorder to fight poverty and poor health conditions. So as a data analyst my role was to categorise the countries based on the clustering techniques and in turn prioritise those countries which are in dire need of this financial aid.

Solution Methodology: For this, the data was first analysed. The outliers in the variables were capped by visualising each variables using the plots and also scaled. Hopkins check was performed inorder to see whether the data is suitable for clustering. K-means clustering method was first checked to begin with. With the help of the Silhouette score and Elbow curve it was decided that the optimal value of k should be three. In order to check on the grounds of a country's socio-economic and health conditions, all the countries were assigned to a cluster using the three variables: income, gdpp and child mortality. The scatterplots and barplots among the three driving factors were plotted to find out which cluster of countries are in dire need of the aid. According to the clusters assigned, the list of 10 countries from that cluster were identified. For Hierarchical clustering, the dendrograms for single linkage and complete linkage dendrograms were inspected. Though from the complete linkage dendrogram, number of clusters four seemed good but on further inspection it was found out that one of the cluster had a huge imbalance showing only 3 three countries in the cluster. So choosing the number of clusters as three seemed a better idea keeping in mind the business perspective. Similar steps were performed for visualising which cluster has the lowest gdpp, lowest income and highest child mortality and similarly the list of 10 countries were made according to the cluster. The list came out to be the same. Thus it was confirmed that those were the countries that needed the financial aid the most.

QUESTION 2A: COMPARE AND CONTRAST K-MEANS CLUSTERING AND HIERARCHICAL CLUSTERING

- K-means clustering is used when the number of clusters to begin with is specified in advance, where as Hierarchical clustering does not require the initial number of clusters to be known in advance.

- K-means clustering is used when there is a huge dataset where as hierarchical is used when there is a small dataset. This is because the algorithm of Hierarchical clustering requires to memorise the assignment of each of the clusters from the initial stage when we have "n" datapoints and "n" clusters to begin with.
- In K-means , there is a cost function that needs to be minimised which is observed as: $J = \sum_{i=1}^{n} ||X_i - \mu_{k(i)}||^2 = \sum_{k=1}^{K} \sum_{i \epsilon C_k} ||X_i - \mu_k||^2$ whereas hierarchical doesn't have anything as such.
- When the clusters shows circular pattern, K-means method is more suitable, where as hierarchical method is suitable for non- circular cluster pattern
- K-means is more sensitive to noise than Hierarchical clustering method.
- K-means has a linear time complexity, whereas hierarchical has quadratic time complexity. Hence k-means is time efficient.

QUESTION 2B : BRIEFLY EXPLAIN THE STEPS OF THE K-MEANS CLUSTERING ALGORITHM

We have to follow the following steps inorder to perform the K-means clustering:

- According to the business problem at hand and the business understanding, the number of clusters are chosen inorder to begin with
- If 'k' clusters have been decided, then 'k' cluster centroid points are selected at random.
- The Euclidean distance of all the data points are calculated with respect to these centroid points and whichever is nearest, the points are assigned to that cluster.
- The centroid points of each cluster are then recalculated to be the mean of the distance of all the data points from the centre based on the Euclidean distance.
- As the centroid points change, the other cluster points are reassigned to new clusters according to those updated centroid points.
- The fourth and the fifth steps are repeated till we get the stable centroid points i.e they stop moving.

QUESTION 2C: HOW IS THE VALUE OF 'K' CHOSEN IN K-MEANS CLUSTERING? EXPLAIN BOTH THE STATISTICAL AS WELL AS THE BUSINESS ASPECT OF IT

For the K-means clustering, it is very important to choose the value of "k". This can be done by using a combination the Silhouette score and elbow curve method. Using these two methods we can easily be able to find out the optimal value of the "k".

In order to get to know about Elbow curve method in a greater detail, we need to know about SSE(sum of squared error) which is basically the sum of the squared distance between each of the data points of a cluster and its centroid point. So mathematically it is: $SSE = \sum_{i-1}^{K} \sum_{x \in c_i} dist(x, c_i)^2$

where the K value represents the number of clusters.

If the values of K is plotted against the SSE values calculated, it can be observed that as the K increases, the error decreases. So the K value is chosen in such a manner where the SSE value falls abruptly. But this method may not give the optimum results. so this method is combined with the Silhouette score method to determine the actual optimal value of the K.

Silhouette coefficient basically signifies the measure of how much a data point in a dataset is similar to its own cluster i.e. cohesion in comparison to the other clusters i.e. separation. Silhouette Score (S(i)) score is determined by using the formula: S(i) = b(i) – a(i)/max{b(i), a(i)} where a(i) gives the average distance from its own cluster and b(i) gives the average distance from the nearest neighbouring cluster. The greater the value of the Silhouette score over the different values of 'K', the more proper is the outcome. But, it is not possible to select the 'K' with the highest silhouette score. For that it is important for us to also understand the business at hand. If there are a lot many clusters, then it will be difficult to understand the business. It is better to take two or three "K" values and check which of them helps in interpreting the business in a more suitable way and also achieve the goal of the business. In this way, we can get the optimal value of the number of clusters "K".

QUESTION 2D: EXPLAIN THE NECESSITY FOR SCALING/STANDARDISATION BEFORE PERFORMING CLUSTERING

standardisation process is generally followed to rescale the values of the variables of our dataset so that they have a common scale among them. In our dataset, each of the variables may have a different scale or each of the variables have a different units. In clustering the data points are grouped based on their distance between the points so there must be some standard in order to produce proper clustering. If each of the variables ion a dataset do not follow a proper standard, then they are not comparable to each other. If the variables with higher scales are standardized, then the forming of clusters won't get affected. Standardization helps to make the relative weight of each variable equal by converting each variable to a unitless measure or relative distance.

QUESTION 2E: EXPLAIN THE DIFFERENT LINKAGES USED IN HIERARCHICAL CLUSTERING

The different linkages used in hierarchical clustering are:

- Single Linkage: In this, the distance between 2 clusters is defined as the shortest distance between points in the two clusters.
- Complete Linkage: In this, the distance between 2 clusters is defined as the maximum distance between any 2 points in the cluster.
- Average Linkage: In this, 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.