

## Assignment Summary

In this assignment we were supposed to give the list of countries that are in direst need of support financially so that the **HELP International's CEO** can help in viewing the values in a bigger picture.

So, we started with the normal EDA process required to analyse and get familiar with data. From the data dictionary we found that the variables **exports, imports & health** were in the percentage of the **GDP**, hence converted these in their absolute values. Then we checked the outliers as they can affect the formation of groups.

On analysing, we found that there are some outliers present, but they seemed to be valid and not statistically insignificant because there we no hypothetical value. So, we kept those and verified the correlation, but that did not give much insight because the data was skewed so we transformed data to make it less skewed.

After transformation, the correlation became more meaningful and the outlier also reduced so the data was ready for the clustering.

We started with the k-means algo and first thing was to get the optimal number if clusters. **For that we used elbow method and Silhouette score and number of clusters came to be 3.**

Then we used hierarchical clustering, we used ward method to get the dendrogram and on analysing that we decided to use 3 groups.

**For both we classified different countries into 3 groups (0 – Under-Developed, 1 – Developing, 2 – Developed) and the group 0 was the one that required attention.**

Between these 2, hierarchical clustering gave better result because the data was pretty well grouped (exports were least, gdpp was low & child\_mort were high).

# Clustering Questions

**Question 1.** Compare and contrast K-means Clustering and Hierarchical Clustering.

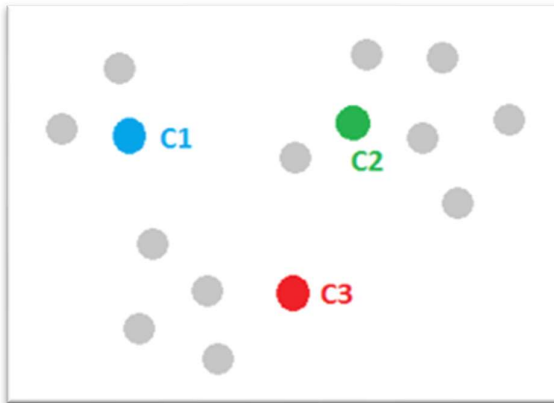
**Answer 1.**

k-means Clustering	Hierarchical Clustering
Requires number of clusters before making model.	No pre knowledge of number of clusters required.
Elbow method or average 2 Average silhouette Method used to get the number of clusters.	Dendrogram used to finalize the number of cluster using the distance threshold.
<ol style="list-style-type: none"><li>1. Initialise some random centres</li><li>2. Assign the datapoint to it based on least distance between centres,</li><li>3. Recalculate the CenterPoint's by mean of points of clusters and repeat step 2 and 3 until no change in the centre points</li></ol>	<ol style="list-style-type: none"><li>Calculate the NxN distance (similarity) matrix.</li><li>2. assigning each item to its own cluster to have N clusters.</li><li>3. Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now you have one less cluster.</li><li>4. Compute distances (similarities) between the new cluster and each of the old clusters.</li><li>5. Repeat steps 3 and 4 until all items are clustered into a single cluster of size N.</li></ol>
It can handle big data as the time complexity of K-means is linear ( $O(n)$ ).	It can't handle big data well as the time complexity of Hierarchical clustering is quadratic ( $O(n^2)$ ).
Since we start with random choice of clusters, the results produced by running the algorithm multiple times might differ. Have to use random_state to keep the result similar	Same results are reproducible

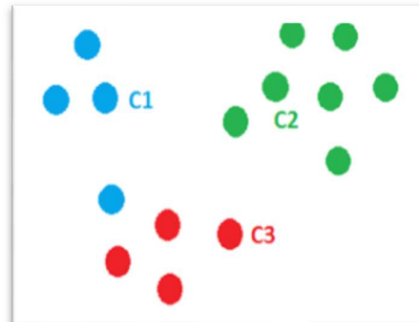
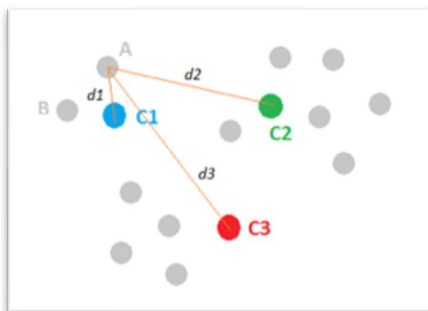
**Question 2.** Briefly explain the steps of the K-means clustering algorithm.

**Answer 2.**

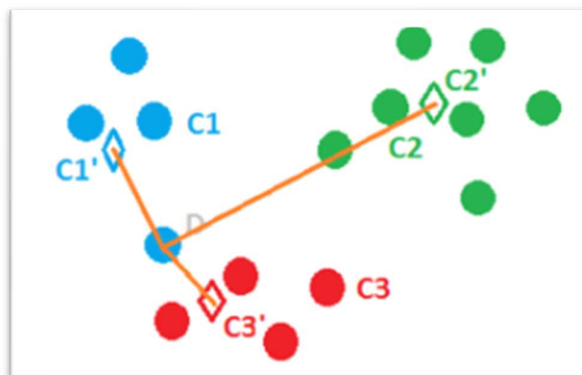
- Initialise some random centres

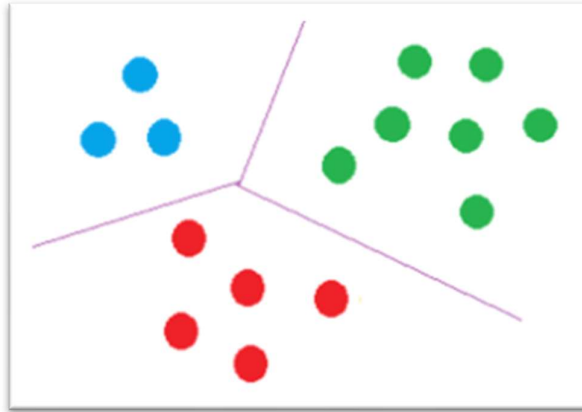


- Assign the datapoint to it based on least distance between centres



- Recalculate the CenterPoint's by mean of points of clusters and repeat step 2 and 3 until no change in the centre points





Result when centres are not changing

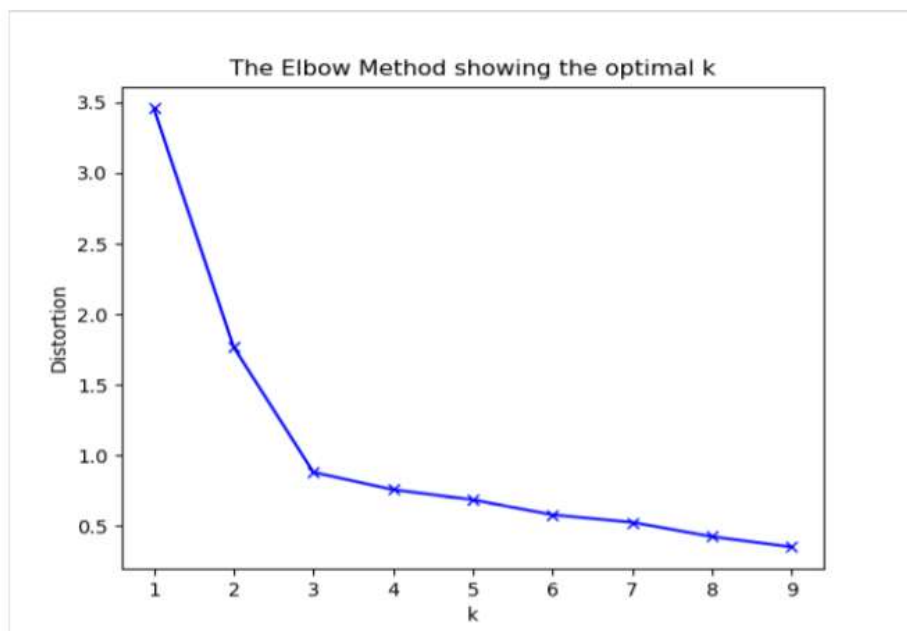
---

**Question 3.** How is the value of 'k' chosen in K-means clustering? Explain both the statistical as well as the business aspect of it.

**Answer 3.** The numbers of clusters can be obtained by following methods:

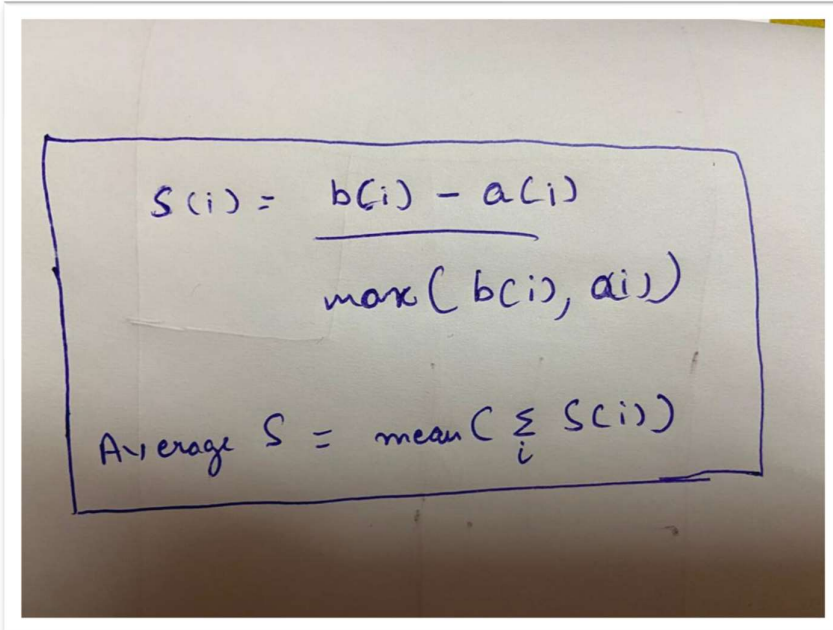
**1. Elbow Method**

In this, the variance (within-cluster sum of squared errors) is plotted against the number of clusters. The first few clusters will introduce a lot of variance and information, but at some point, the information gain becomes low, thus creating an elbow shape. The optimal number of clusters is found at this kink



## 2. Average silhouette Method

In this Silhouette value for every datapoint is calculated, the mean of which is used to find the optimal number of clusters. The silhouette value represents how similar a datapoint is to its own cluster when compared to all the other clusters or cluster centroids. The value ranges from -1 to +1. A higher silhouette value implies that the datapoint is matched well to its own centroid/cluster and is not so well matched with other clusters. If the mean of the silhouette value measured for all the datapoints is considerably high, then it can be said that the number of clusters are at its optimal value, or in other words, the clustering structure is appropriate. On the other hand, if the mean silhouette value turns out to be very less or negative, then it means that the cluster structure is not proper, and it may be having either more or lesser number of clusters than the optimal value.



The image shows a handwritten formula for the Average Silhouette Method, enclosed in a hand-drawn rectangular box. The formula for the silhouette value  $S(i)$  is written as:

$$S(i) = \frac{b(i) - a(i)}{\max(b(i), a(i))}$$

Below this, the formula for the Average Silhouette is written as:

$$\text{Average } S = \text{mean}\left(\sum_i S(i)\right)$$

But sometimes we need to consider the business requirement also and then choose the number of clusters. For example, if using all these methods we found that the optima number of clusters came to be 5 but business does not require or wants more than 3 clusters then we need to accept that and use 3 and the total number of clusters.

**Question 4.** Explain the necessity for scaling/standardisation before performing Clustering.

**Answer 4.** Standardisation of data, that is, converting them into z-scores with mean 0 and standard deviation 1, is important for 2 reasons in K-Means algorithm:

- 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.

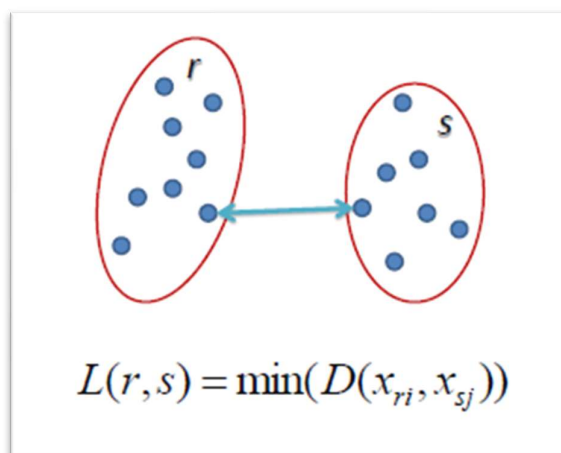
**Example:** When you are working with data where each variable means something different, (e.g., year, income & weight) the fields are not directly comparable. One year is not equivalent to one kg, and may or may not have the same level of importance in sorting a group of records. In this situation where one field has a much greater range of value than another, it may end up being the primary driver of what defines clusters. Standardisation helps to make the relative weight of each variable equal by converting each variable to a unit-less measure or relative distance.

---

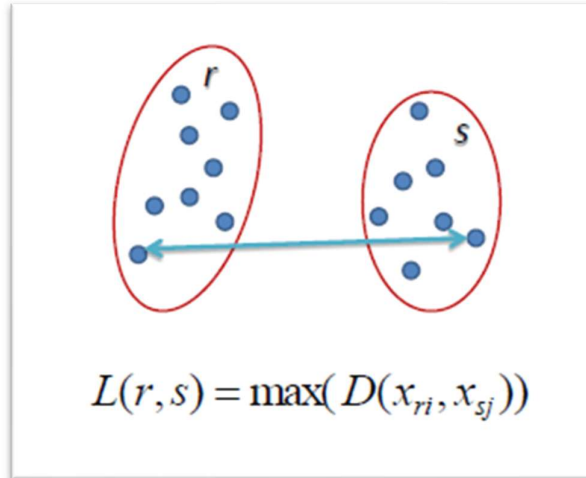
**Question 5.** Explain the different linkages used in Hierarchical Clustering.

**Answer 5.**

1. **Single Linkage:** For two clusters  $r$  and  $s$ , this returns the minimum distance between two points  $i$  and  $j$  such that  $i$  belongs to  $A$  and  $j$  belongs to  $B$ .



2. **Complete Linkage:** For two clusters  $r$  and  $s$ , the single linkage returns the maximum distance between two points  $i$  and  $j$  such that  $i$  belongs to  $r$  and  $j$  belongs to  $s$ .



3. **Average Linkage:** For two clusters  $r$  and  $s$ , first for the distance between any data-point  $i$  in  $r$  and any data-point  $j$  in  $s$  and then the arithmetic mean of these distances are calculated. Average Linkage returns this value of the arithmetic mean.

