Question 1: Assignment Summary

Briefly describe the "Clustering of Countries" assignment that you just completed within 200-300 words. Mention the problem statement and the solution methodology that you followed to arrive at the final list of countries. Explain your main choices briefly (what EDA you performed, which type of Clustering produced a better result and so on)

Note: You don't have to include any images, equations or graphs for this question. Just text should be enough.

Answer:

HELP International is an international humanitarian 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. HELP have been able to raise $ 10 million in the recent funding programmes. CEO of the NGO needs to decide how to use raised money strategically and effectively so that the countries that are in the direst need of aid would get the first priority from organization.

Problem Statement:

As a Data Analyst, needs to categorise the countries using some socio-economic and health factors that determine the overall development of the country. Then analyst need to suggest the countries which the CEO needs to focus on the most. Make sure that analyst report back at least 5 countries which are in direst need of aid from the analysis.

Solution Methodology:

Firstly, our task is to cluster the countries by the factors mentioned above. To do so, we would need to clean and prepare the raw data of countries with socio-economic and health factors. Perform clustering on data and analysing the formed clusters by comparing how these three variables - [gdpp, child_mort and income] vary for each cluster of countries to recognise and differentiate the clusters of developed countries from the clusters of under-developed countries.

We have performed EDA task and created histograms to compare the distribution of a variable across levels and also plotted boxplot to identify the outliers and capped them with 99 percentiles to treat the outliers in data.

We have performed both the types of clustering – K-means and Hierarchical. But the hierarchical clustering is more understandable than K-means and also easy to perform clustering based on countries because we don't need to assume K-value.

Final list of countries:

1.	Liberia	2.	Burundi
3.	Congo, Dem. Rep.	4.	Niger
5.	Sierra Leone		

Question 2: Clustering

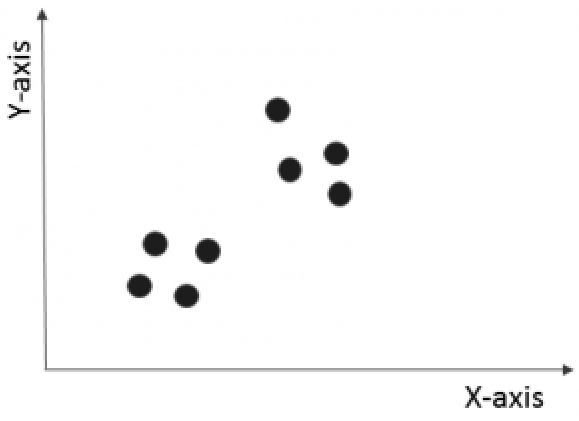
a) Compare and contrast K-means Clustering and Hierarchical Clustering.

Answer:

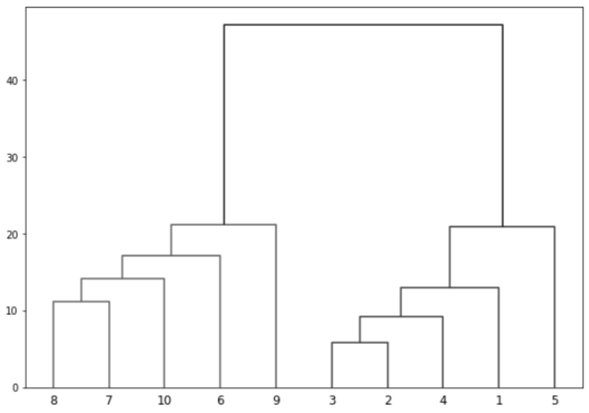
I would say hierarchical clustering is usually preferable, as it is both more flexible and has fewer hidden assumptions about the distribution of the underlying data.

With k-Means clustering, you need to have a sense ahead-of-time of what your desired number of clusters is ('k' value). Also, k-means will often give unintuitive results if your data is not well-separated into sphere-like clusters, or you pick a 'k' not well-suited to the shape of your data, i.e. you pick a value too high or too low, or you have weird initial values for your cluster centroids (one strategy is to run a bunch of k-means algorithms with random starting centroids and take some common clustering result as the final result).

In contrast, hierarchical clustering has fewer assumptions about the distribution of your data - the only requirement (which k-means also shares) is that a distance can be calculated each pair of data points. Hierarchical clustering typically 'joins' nearby points into a cluster, and then successively adds nearby points to the nearest group. You end up with a 'dendrogram', or a sort of connectivity plot. You can use that plot to decide how many clusters your data has, by cutting the dendrogram at different heights. Of course, if you need to pre-decide how many clusters you want (based on some sort of business need) you can do that too. Hierarchical clustering can be more computationally expensive but usually produces more intuitive results.



K-Means Clustering



Hierarchical Clustering

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

Answer:

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

At this point, we arrive at the optimal clusters.

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 K-Means algorithm is simple and perhaps the most commonly used algorithm for clustering.

The basic idea behind k-means consists of defining k clusters such that total within-cluster variation (or error) is minimum.

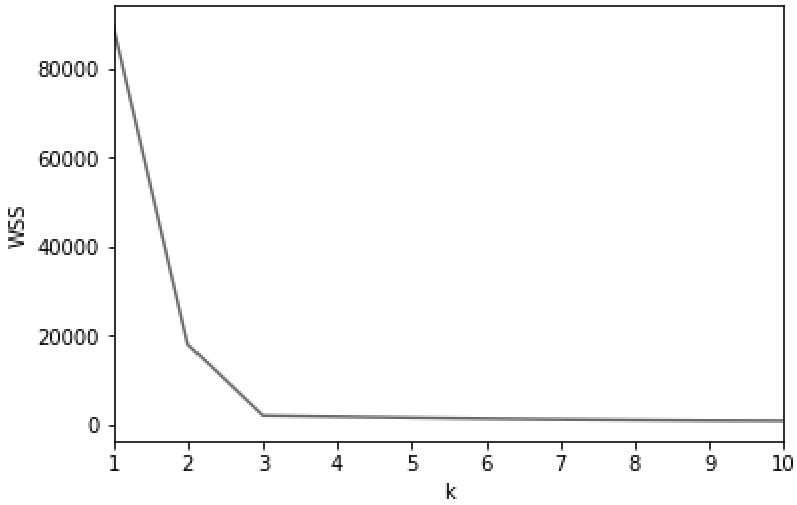
There are several statistical measures available for selecting K. These measures are often applied in com-bination with probabilistic clustering approaches. They are calculated with certain assumptions about the underlying distribution of the data.

Two methods that can be useful to find this mysterious k in k-Means. These methods are:

1. The Elbow Method
2. The Silhouette Method

The Elbow Method

Calculate the Within-Cluster-Sum of Squared Errors (WSS) for different values of k, and choose the k for which WSS becomes first starts to diminish. In the plot of WSS-versus-k, this is visible as an elbow.



As expected, the plot looks like an arm with a clear elbow at k = 3. Unfortunately, we do not always have such clearly clustered data. This means that the elbow may not be clear and sharp. In such an ambiguous case, we may use the Silhouette Method.

The Silhouette Method

The silhouette value measures how similar a point is to its own cluster (cohesion) compared to other clusters (separation).

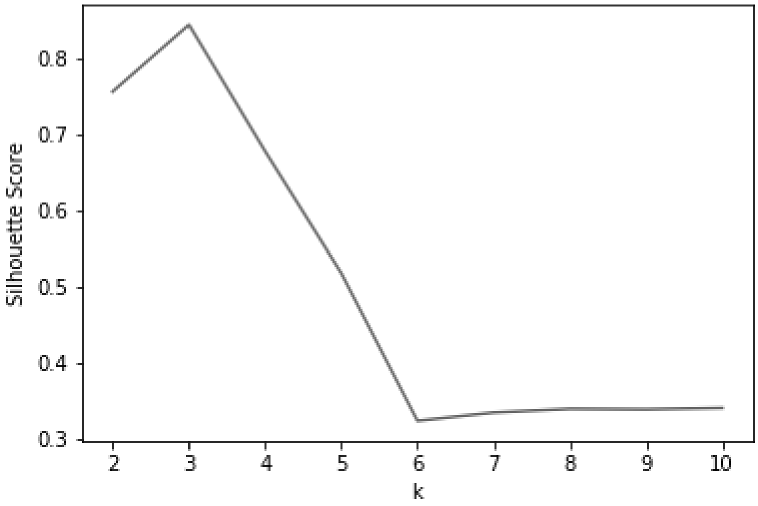
The range of the Silhouette value is between +1 and -1. A high value is desirable and indicates that the point is placed in the correct cluster. If many points have a negative Silhouette value, it may indicate that we have created too many or too few clusters.

The Silhouette Value s(i) for each data point i is defined as follows:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \text{ if } |C_i| > 1$$

and

$$s(i) = 0, \text{ if } |C_i| = 1$$



There is a clear peak at k = 3. Hence, it is optimal. Finally, the data can be optimally clustered into 3 clusters.

The Elbow Method is more of a decision rule, while the Silhouette is a metric used for validation while clustering. Thus, it can be used in combination with the Elbow Method.

Therefore, the Elbow Method and the Silhouette Method are not alternatives to each other for finding the optimal K. Rather they are tools to be used together for a more confident decision.

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

Answer:

Clustering algorithms are certainly affected by the feature scaling.

Example:

Let's say that you have two features:

1. weight (in Lbs)
2. height (in Feet)

... and we are using these to predict whether a person needs a 'S' or 'L' size shirt.

We are using weight + height for that, and in our trained set let's say we have two people already in clusters:

1. Adam (175Lbs+5.9ft) in 'L'
2. Lucy (115Lbs+5.2ft) in 'S'.

We have a new person - Alan (140Lbs+6.1ft.), and your clustering algo will put it in the cluster which is nearest. So, if we don't scale the features here, the height is not having much effect and Alan will be allotted in 'S' cluster.

So, we need to scale it. Scikit Learn provides many functions for scaling. Scaling affects Clustering Results in a way that depends by the metric used (Euclidean Distance, Squared Euclidean Distance, Manhattan Distance, …)

Clustering algorithms such as K-means do need feature scaling before they are fed to the algo. Since, clustering techniques use **Euclidean Distance** to form the cohorts, it will be wise e.g. to scale the variables having heights in meters and weights in KGs before calculating the distance.

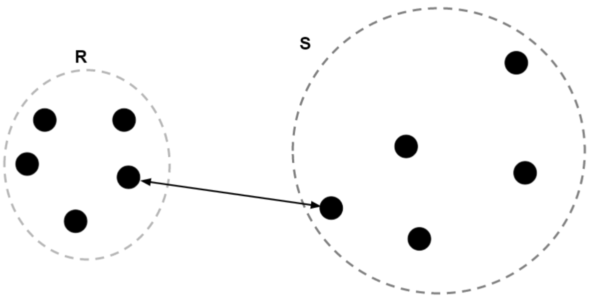
e) Explain the different linkages used in Hierarchical Clustering.

Answer:

The different types of linkages describe the different approaches to measure the distance between two sub-clusters of data points. The different types of linkages are: -

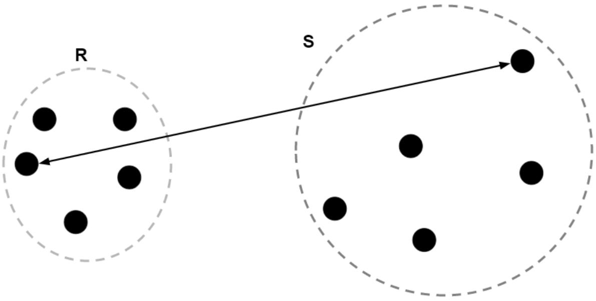
Single Linkage: For two clusters R and S, the single linkage returns the minimum distance between two points i and j such that i belongs to R and j belongs to S.

L(R,S) = min(D(i,j)), i ∑ R, j ∑ S



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.

L(R,S) = max(D(i,j)), i ∑ R, j ∑ S



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.

$$L(R,S) = \frac{1}{n_R+n_S} \sum_{i=1}^{n_R} \sum_{j=1}^{n_S} D(i,j), i\epsilon R, j\epsilon S$$

where

– Number of data-points in R

– Number of data-points in S