Q. 1) Assignment Summary. Briefly describe the "Clustering of countries" assignment.

CLUSTERING ASSIGNMENT

Q. 1. Assignment Summary

Ans → The objectives of this assignment is to form the clusters of countries using socio-economic and health factors that determine the overall development of the country and suggests the countries which the HELP NGO international CEO needs to focus on the most.

Start off with the necessary data inspection i.e. size of datasets, check the NULL percentage of rows and columns, etc. and then EDA activities on the provided dataset. For example - converts exports, imports and health percentage to the absolute value of their GDP per capita, univariate & bivariate analysis, etc.

Performed outlier analysis on the dataset. For this we have flexibility of not removing the outliers if it suits the business needs or lot of countries will get removed. Hence then just find the outliers in the dataset. For outlier treatment, done the percentile capping with IQR method to handle the outliers in the dataset.

Checked the Hopkins statistics test on the dataset to measure the cluster tendency of the dataset. Ran the kernel multiple time to calculate Hopkins score to know whether the data is good for clustering or not and got score always $\geq 80\%$.

Scaled the dataset before applying models with standardisation scaling method.
Then, applied both the clustering algorithm K-Means and Hierarchical clustering on the dataset.

In the K-Means, need to find optimal no. of clusters for that performed sum of squared (Elbow curve) method and then performed Silhouette analysis. From this 2 method, found the optimal value of K (K=3) clusters into which data clustered.

From the business understanding we know that Child_mortality, Income and GDPP are important factors which decides the development of country. Based on these 3 factors formed the cluster and got the countries which we need to focus on.

Last, performed the Hierarchical clustering with single linkage and complete linkage. Performed cluster profiling on the data and identify the countries which are in dire need of aid. Also performed visualisation on cluster that have been formed and identified countries.

Finally we got the same list of countries from both K-Means and Hierarchical clustering which needs to be focused for socio-economic and health growth.

Q. 2) a) Compare and contrast K-Means Clustering and Hierarchical Clustering.

Ans – K-Means clustering is unsupervised learning type of algorithm in machine learning. It's the process of dividing N data points into k no. of clusters.

Hierarchical clustering is also unsupervised learning type of algorithm in machine learning. As the name suggests, is an algorithm that builds hierarchy of clusters. This algorithm starts with all the data points assigned to a cluster of their own. Then two nearest clusters are merged into the same cluster. In the end, this algorithm terminates when there is only a single cluster left.

	K – Means Clustering	**Hierarchical Clustering**
1	K-Means clustering, using a pre-specified number of clusters, the method assigns records to each cluster to find the mutually exclusive cluster of spherical shape based on distance.	Hierarchical methods can be either divisive or agglomerative.
2	K-Means clustering needed advance knowledge of K i.e. no. of clusters one wants to divide your data.	In hierarchical clustering one can stop at any number of clusters, one finds appropriate by interpreting the dendrogram.
3	K- means clustering a simply a division of the set of data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset).	A hierarchical clustering is a set of nested clusters that are arranged as a tree.
4	One can use median or mean as a cluster centre to represent each cluster.	Agglomerative methods begin with 'n' clusters and sequentially combine similar clusters until only one cluster is obtained.
5	Methods used are normally less computationally intensive and are suited with very large datasets.	Divisive methods work in the opposite direction, beginning with one cluster that includes all the records and Hierarchical methods are especially useful when the target is to arrange the clusters into a natural hierarchy.
6	Advantage – Fast, robust and easier to understand.	Advantage - Ease of handling of any forms of similarity or distance.

Q. 2) 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. Among all the unsupervised learning algorithms, clustering via k-means might be one of the simplest and most widely used algorithms. Briefly speaking, k-means clustering aims to find the set of k clusters such that every data point is assigned to the closest centre, and the sum of the distances of all such assignments is minimized. Here the steps of the algorithm are:

a) Start by choosing K random points the initial cluster centres.

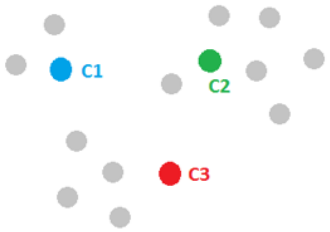
b) Assign each data point to their nearest cluster centre. The most common way of measuring the distance between the points is the Euclidean distance.
c) For each cluster, compute the new cluster centre which will be the mean of all cluster members.
d) Now re-assign all the data points to the different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

For Example

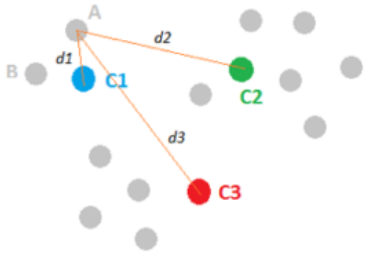
Let's walk through a simple example to better understand the idea. Imaging we have these grey points in the following figure and want to assign them into three clusters. K-means follows the four steps listed below.

Step one: Initialize cluster centre's

We randomly pick three points C1, C2 and C3, and label them with blue, green and red colour separately to represent the cluster centres.

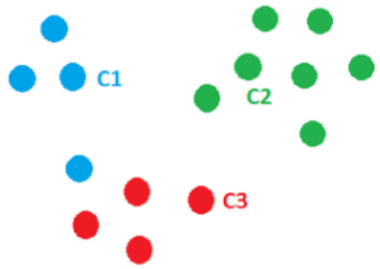


Step two: Assign observations to the closest cluster centre.



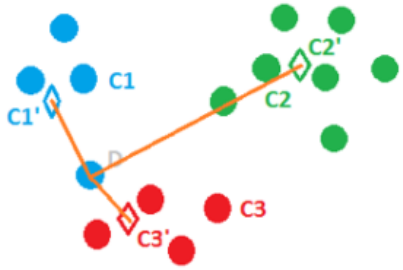
Once we have these cluster centre's, we can assign each point to the clusters based on the minimum distance to the cluster centre. For the grey point A, compute its distance to C1, C2 and C3, respectively. And after comparing the lengths of *d1*, *d2* and *d3*, we figure out

that *d1* is the smallest, therefore, we assign point A to the blue cluster and label it with blue. We then move to point B and follow the same procedure. This process can assign all the points and leads to the following figure.



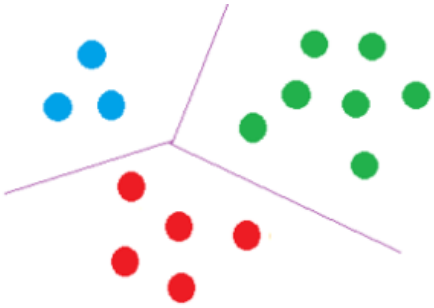
Step three: Revise cluster centre's as mean of assigned observations

Now we've assigned all the points based on which cluster centre they were closest to. Next, we need to update the cluster centres based on the points assigned to them. For instance, we can find the centre mass of the blue cluster by summing over all the blue points and dividing by the total number of points, which is four here. And the resulted centre mass C1', represented by a blue diamond, is our new centre for the blue cluster. Similarly, we can find the new centres C2' and C3' for the green and red clusters.



Step four: Repeat step 2 and step 3 until convergence

The last step of k-means is just to repeat the above two steps. For example, in this case, once C1', C2' and C3' are assigned as the new cluster centres, point D becomes closer to C3' and thus can be assigned to the red cluster. We keep on iterating between assigning points to cluster centres, and updating the cluster centre's until convergence. Finally, we may get a solution like the following figure.

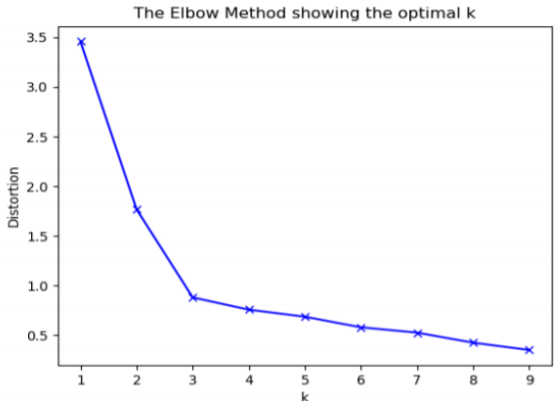


Final cluster would be looks like above.

Q. 2) 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 - The value of 'k' chosen in K-Means clustering by various methods but most common methods used in industry are Elbow method and Average Silhouette Method.

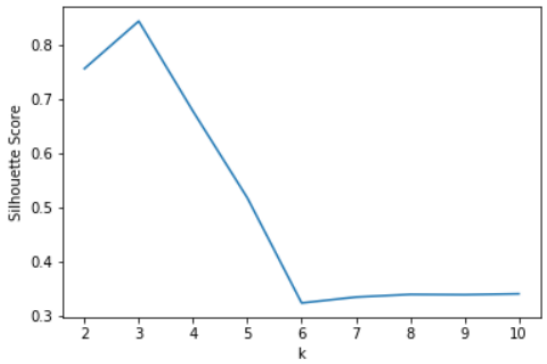
a) **Elbow Method** – The Elbow Method is one of the most popular methods to determine this optimal value of k.
Compute clustering algorithm (e.g., k-means clustering) for different values of k. For instance, by varying k from 1 to 10 clusters.
For each k, calculate the total within-cluster sum of square (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.



b) **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 silhouette).
Plot the curve of avg.silhouette according to the number of clusters k.
The location of the maximum is considered as the appropriate number of clusters.

Silhouette Score=(p−q)/max(p,q)

p is the mean distance to the points in the nearest cluster that the data point is not a part of
q is the mean intra-cluster distance to all the points in its own cluster.
The value of the silhouette score range lies between -1 to 1.
A score closer to 1 indicates that the data point is very similar to other data points in the cluster,
A score closer to -1 indicates that the data point is not similar to the data points in its cluster. Below is the figure of Silhouette Score.



Q. 2) d) Explain the necessity for scaling/standardisation before performing clustering.

Ans - Standardization (also called data normalization or feature scaling) refers to the process of rescaling the values of the variables in your data set so they share a common scale. Often performed as a pre-processing step, particularly for cluster analysis, standardization may be important if you are working with data where each variable has a different unit (e.g., inches, meters, tons and kilograms), or where the scales of each of your variables are very different from one another (e.g., 0-1 vs 0-1000). The reason this importance is particularly high in cluster analysis is because groups are defined based on the distance between points in mathematical space.

For example - Suppose we have 3 variables – Income, Number of cars owned and Age. Clearly these 3 variables are measuring very different things, and thus have very different scales. If we perform cluster analysis on this data, differences in income will most likely dominate the other 2 variables simply because of the scale. In most practical cases, all these different variables need to be converted to one scale in order to perform meaningful analysis.

Standardization prevents variables with larger scales from dominating how clusters are defined. It allows all variables to be considered by the algorithm with equal importance.

There are a few different options for standardization, but two of the most frequently used are z-score and unit interval:

1. Z-score transforms data by subtracting the mean value for each field from the values of the file and then dividing by the standard deviation of the field, resulting in data with a mean of zero and a standard deviation of one.
2. Unit interval is calculated by subtracting the minimum value of the field and then dividing by the range of the field (maximum minus minimum) which results in a field with values ranging from 0 to 1.

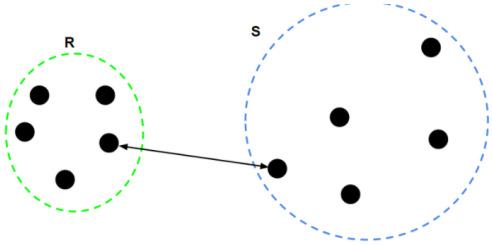
Although standardization is considered best practice for cluster analysis.

Q. 2) e) Explain different linkage used in Hierarchical clustering.

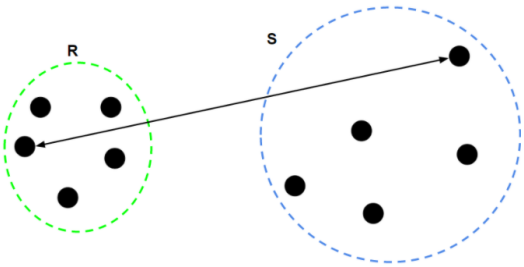
Ans - Hierarchical Clustering involves either clustering sub-clusters (data points in the first iteration) into larger clusters in a bottom-up manner (which is called Agglomerative approach) or dividing a larger cluster into smaller sub-clusters in a top-down manner (which is called

Divisive approach). During both the types of hierarchical clustering, the distance between two sub-clusters needs to be computed. 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 –

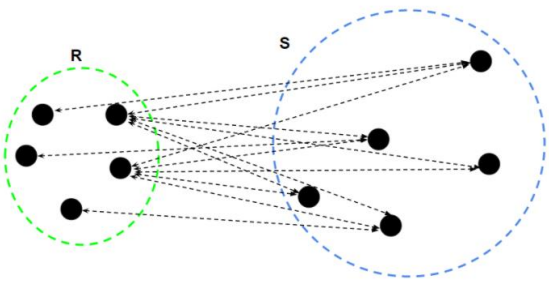
a) **Single Linkage** - Single-linkage (nearest neighbour) is the shortest distance between a pair of observations in two clusters. It can sometimes produce clusters where observations in different clusters are closer together than to observations within their own clusters.
For two clusters R and S, the single linkage returns the minimum distance between two points i and j such that i belongs to R and j belongs to S.



b) **Complete Linkage –** Complete-linkage (farthest neighbour) is where distance is measured between the farthest pair of observations in two clusters. This method usually produces tighter clusters than single-linkage, but these tight clusters can end up very close together. Along with average-linkage, it is one of the more popular distance metrics.
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.



c) **Average Linkage –** Average-linkage is where the distance between each pair of observations in each cluster are added up and divided by the number of pairs to get an average inter-cluster distance. Average-linkage and complete-linkage are the two most popular distance metrics in hierarchical clustering.
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.



d) **Centroid Linkage -** Centroid-linkage is the distance between the centroids of two clusters. As the centroids move with new observations, it is possible that the smaller clusters are more similar to the new larger cluster than to their individual clusters causing an inversion in the dendrogram. This problem doesn't arise in the other linkage methods because the clusters being merged will always be more similar to themselves than to the new larger cluster.

Submitted By
Pritamkumar Suryavanshi