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)

Answer-

1. Import necessary libraries.
2. Reading and cleaning the data.
3. Data visualization
4. Checking outlier
5. Clustering

    a. Hopkins check

    i. The Hopkins statistic, is a statistic which gives a value which indicates the cluster tendency, in other words: how well the data can be clustered.

    ii. We can see that the value is between {0.8, ..., 0.99}, so the dataset has a high tendency to cluster.

6. K-Mean Clustering

    a. To Choose the value of k, there are two methods, 1. Silhouette score 2. Elbow curve-ssd

    i. Silhouette score
silhouette score=$p-qmax(p,q)$silhouette score=p−qmax(p,q)
pp is the mean distance to the points in the nearest cluster that the data point is not a part of
qq 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.

    ii. Elbow curve

7. Clustering profiling

    a. From cluster profiling in K- means clustering we can see that :

    b. 1. Cluster 0 is having the High income, High GDP and very Low child mortality

    c. 2. Cluster 2 is having very Low income, very Low GDP but High child mortality

    d. 3. Cluster 1 is having low income, GDP and less child mortality

8. Hierarchical Clustering

    a. Hierarchical clustering involves creating clusters that have a predetermined ordering from top to bottom. For example, all files and folders on the hard disk are organized in a hierarchy. There are two types of hierarchical clustering,

    i. Divisive

ii. Agglomerative.

b. Single Linkage

i. In single linkage hierarchical clustering, the distance between two clusters is defined as the shortest distance between two points in each cluster. For example, the distance between clusters “r” and “s” to the left is equal to the length of the arrow between their two closest points.

c. Complete Linkage

i. In complete linkage hierarchical clustering, the distance between two clusters is defined as the longest distance between two points in each cluster. For example, the distance between clusters “r” and “s” to the left is equal to the length of the arrow between their two furthest points.

d. From cluster profiling using hierarchical clustering we can see that :

i. 1. Cluster 0 is having the High child mortality, low GDP and very Low child mortality

ii. 2. Cluster 1 is having Low child mortality, moderate income and GDP

iii. 3. Cluster 2 is having very low child mortality,high income and GDP

9. Final Analysis

a. From K means clustering we got better clusters compared to Hierarchical clustering.¶
Cluster 2 is the better cluster we got with High child mortality, low income and low GDP
Final list of country we got.

10. Inferences

a. From the EDA performed we could see that Income, GDP and child Mortality are the major three variables need to be focused

b. In K means clustering we got Cluster 2 is having very Low income, very Low GDP but High child mortality. So we concluded that countries under cluster 2 are in need of aid.

c. In Hierarchical clustering we saw that Cluster 0 is having the High child mortality, low GDP and very Low child mortality.

d. The clusters formed in Hierarchical clustering were not that good. So we went on to consider cluster formed in K means clustering. And got top five countries with High child mortality,Low GDP and Low income
Then we looked for the countries based on socio economic factors

11. Recommendations

a. From the analysis performed, We can see that low income people have high child mortality, so CEO must focus more on low income countries

b. We could also see Low GDP per capita countries are not having much import and export of goods and services. Those countries also must be focused

c. There are some countries which spend well on health for the people living in that country. For ex: US. Such countries can be skipped. And focus more on Burundi, Congo, Dem. Rep where the total health spending is too less.

d. If the total fertility is less the life expectency is more. Haiti is the country having very low life expectancy, and high child mortality. Its good to have less children per woman,so that they could be looked after well.¶

Question 2: Clustering

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

a. Ans-

k-means Clustering	**Hierarchical Clustering**
k-means, 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.
K Means clustering needed advance knowledge of K i.e. no. of clusters one want to divide your data.	In hierarchical clustering one can stop at any number of clusters, one find appropriate by interpreting the dendrogram.
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.
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.
In K Means clustering, since one start with random choice of clusters, the results produced by running the algorithm many times may differ.	In Hierarchical Clustering, results are reproducible in Hierarchical clustering
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.
K Means clustering is found to work well when the structure of the clusters is hyper spherical (like circle in 2D, sphere in 3D).	Hierarchical clustering don't work as well as, k means when the shape of the clusters is hyper spherical.
Advantages:	Advantages:

1. Convergence is guaranteed. 2. Specialized to clusters of different sizes and shapes.	1. Ease of handling of any forms of similarity or distance. 2. Consequently, applicability to any attribute’s types.
Disadvantages: 1. K-Value is difficult to predict 2. Didn’t work well with global cluster.	Disadvantage: 1. Hierarchical clustering requires the computation and storage of an n×n distance matrix. For very large datasets, this can be expensive and slow

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

Algorithmic steps for k-means clustering

Let X = {x_1,x_2,x_3,……..,x_n} be the set of data points and V = {v_1,v_2,…….,v_c} be the set of centers.

1) Randomly select ‘c’ 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 of all the cluster centers.

4) Recalculate the new cluster center using:
where, ‘ci’ represents the number of data points in ith cluster.

$$v_i = (1/c_i) \sum_{j=1}^{c_i} x_i$$

5) Recalculate the distance between each data point and new obtained cluster centers.

6) If no data point was reassigned then stop, otherwise repeat from step 3).

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- Determining the optimal number of clusters in a data set is a fundamental issue in partitioning clustering, such as k-means clustering, which requires the user to specify the number of clusters k to be generated.

Unfortunately, there is no definitive answer to this question. The optimal number of clusters is somehow subjective and depends on the method used for measuring similarities and the parameters used for partitioning.

A simple and popular solution consists of inspecting the dendrogram produced using hierarchical clustering to see if it suggests a particular number of clusters. Unfortunately, this approach is also subjective.

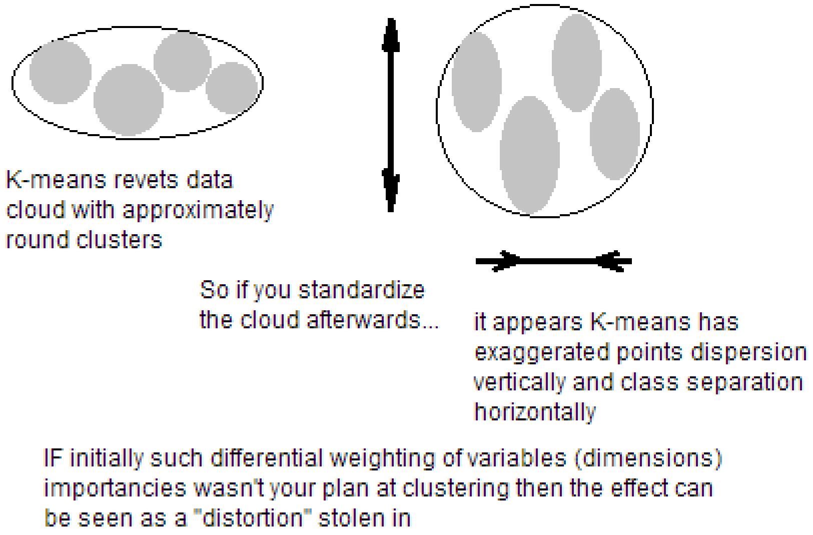
In this chapter, we'll describe different methods for determining the optimal number of clusters for k-means, k-medoids (PAM) and hierarchical clustering.

These methods include direct methods and statistical testing methods:

1. Direct methods: consists of optimizing a criterion, such as the within cluster sums of squares or the average silhouette. The corresponding methods are named elbow and silhouette methods, respectively.
2. Statistical testing methods: consists of comparing evidence against null hypothesis. An example is the gap statistic.

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

If your variables are of incomparable units (e.g. height in cm and weight in kg) then you should standardize variables, of course. Even if variables are of the same units but show quite different variances it is still a good idea to standardize before K-means. You see, K-means clustering is "isotropic" in all directions of space and therefore tends to produce more or less round (rather than elongated) clusters. In this situation leaving variances unequal is equivalent to putting more weight on variables with smaller variance, so clusters will tend to be separated along variables with greater variance.



K-means revets data cloud with approximately round clusters

So if you standardize the cloud afterwards...

it appears K-means has exaggerated points dispersion vertically and class separation horizontally

IF initially such differential weighting of variables (dimensions) importancies wasn't your plan at clustering then the effect can be seen as a "distortion" stolen in

A different thing also worth to remind is that K-means clustering results are potentially sensitive to the order of objects in the data set11. A justified practice would be to run the analysis several times, randomizing objects order; then average the cluster centres of those runs and input the centres as initial ones for one final run of the analysis.

e) *Explain the different linkages used in Hierarchical Clustering.*

a. Single Linkage
    i. In single linkage hierarchical clustering, the distance between two clusters is defined as the shortest distance between two points in each cluster. For example, the distance between clusters “r” and “s” to the left is equal to the length of the arrow between their two closest points.

b. Complete Linkage
    ii. In complete linkage hierarchical clustering, the distance between two clusters is defined as the longest distance between two points in each cluster. For example, the distance between clusters “r” and “s” to the left is equal to the length of the arrow between their two furthest points.