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).

Solution:

The objective of the assignment is to find the top countries who are in the direst need of Aid based on the Socio-Economic and Health Factors which determine the overall development of the society. The following approach has been used by me to reach the objective.

I have used the unsupervised clustering technique to identify a pattern among the countries and to provide an insight in to the natural groupings found within the data.

Step 1: Reading and Understanding the Data

The data set provided was read into the juypter notebook and the initial statistical parameters were determined.

Step2: Data Cleaning

This is a critical stage, in the he missing values and the null values are checked. Here there were no missing values were found in the data set.

Step 3: Bi- Variate Analysis & Data Visualization

In this step the countries have been studied w.r.t the gdpp, income and child_mort to understand which countries vary with the variables

The correlation among the continuous numerical values using correlation matrix and heat map was also observed.

Step 4: Data Visualization

The continuous values have been plotted using pair plots to understand the pattern behaviour among the variables.

Step 5: Data Preparation (Data Metrics) and visualizing distribution of the data for cluster formations

The three variables which are given in the data set have been transformed to their original values and distplots have been plotted to visualize distribution of the data and cluster formations.

Step 6: Outlier Treatment

Visualizing the columns for outlier treatment. Have performed soft capping on the data set, as the dataset is very small and wouldn't like to drop the rows.

Step 7: Hopkins Statistics Test

The Hopkins statistic (introduced by Brian Hopkins and John Gordon Skellam) is a way of measuring the cluster tendency of a data set. It acts as a statistical hypothesis test where the null and the alternative hypotheses are defined as follow:

- **Null hypothesis**: the data set is uniformly distributed (i.e., no meaningful clusters)

- **Alternative hypothesis**: the data set is not uniformly distributed (i.e., contains meaningful clusters)

A value close to 1 tends to indicate the data is highly clustered, random data will tend to result in values around 0.5, and uniformly distributed data will tend to result in values close to 0.
Put in other words, If the value of Hopkins statistic is close to 1, then we can reject the null hypothesis and conclude that the dataset is significantly a clusterable data.
In our assignment we get a score of 80% and above.

Step 8: Model Building
It is extremely important to rescale the variables so that they have a comparable scale. There are two common ways of rescaling:

1. Min-Max scaling
2. Standardisation (mean-0, sigma-1)

In the assignment, we will use Standard Scaling.

Min Max Scaling is a good technique to use when you do not know the distribution of your data or when you know the distribution is not Gaussian (a bell curve). Whereas **Standardization** assumes that your data has a Gaussian (bell curve) distribution.

Step 9: K-Means Clustering
K-means clustering is one of the simplest and popular unsupervised machine learning algorithms.

The algorithm works as follows:

1. First, we initialize k points, called means, randomly.
2. We categorize each item to its closest mean and we update the mean's coordinates, which are the averages of the items categorized in that mean so far.
3. We repeat the process for a given number of iterations and at the end, we have our clusters.

To get the optimal number of Clusters:

1. **Elbow Curve (Min Distance)**

The Elbow Method is one of the most popular methods to determine this optimal value of k.

2. **Silhouette Analysis (Max Distance)**

$$\text{silhouette score} = p - q / max\ (p, q)$$

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.

1. A score closer to 1 indicates that the data point is very similar to other data points in the cluster,

2. A score closer to -1 indicates that the data point is not similar to the data points in its cluster.

In the assignment I have taken the value of K as 3 to proceed further as optimal number of clusters and preferred Hierarchical Clustering as dataset is very less

Step 9: Drawing inferences from the clusters formed

Filtering of cluster formed which is of our requirement high child mortality, low income and low gdpp

S.No	Country
66	Haiti
132	Sierra Leone
32	Chad
31	Central African Republic
97	Mali

Question 2: Clustering

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

Solution:

While carrying on an unsupervised learning task, the data provided with is not labelled. It means that the algorithm will aim at inferring the inner structure present within data, trying to group, or cluster, them into classes depending on similarities among them. There are two main clusterization algorithms:

1. K-means: K means is an iterative clustering algorithm that aims to find local maxima in each iteration
2. Hierarchical clustering: Hierarchical clustering, 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.

Comparison between K Mean Clustering and Hierarchical Clustering

Category	K Means Clustering	Hierarchical Clustering
Algorithm	Centroid Method	Hierarchical / Agglomerative
Method for Optimal no of Clusters	Elbow method	Dendrogram
Directional choice of Clusters	Random	Top-Down/Bottom-up
Data Handling	Big Data can be easily handled	Big Data can’t be handled as it involves lot of calculation
Results	The final clusters might differ due to random initial clusters	The final clusters are constant/ reproducible
Value of K	Prior value of K is required	Prior Value of K is not required, depends on interpretation from dendrogram

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

Solution

The algorithm which tries to minimize the sum of distances between the points in a cluster with their centroid is called as the k-means clustering technique.

K-means is a centroid-based algorithm, or a distance-based algorithm, where we calculate the distances to assign a point to a cluster. In K-Means, each cluster is associated with a centroid.

Let's understand how K-Means actually works:

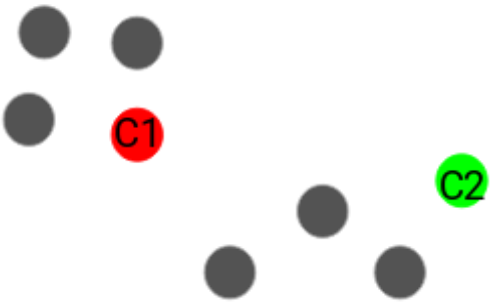
We have these 8 points and we want to apply k-means to create clusters for these points. Here's how we can do it.

Step 1: Choose the number of clusters k

The first step in k-means is to pick the number of clusters, k.

Step 2: Select k random points from the data as centroids

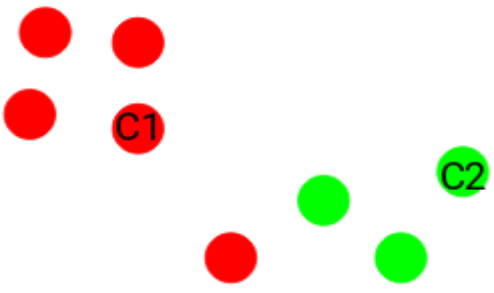
Next, we randomly select the centroid for each cluster. Let's say we want to have 2 clusters, so k is equal to 2 here. We then randomly select the centroid:



Here, the red and green circles represent the centroid for these clusters.

Step 3: Assign all the points to the closest cluster centroid

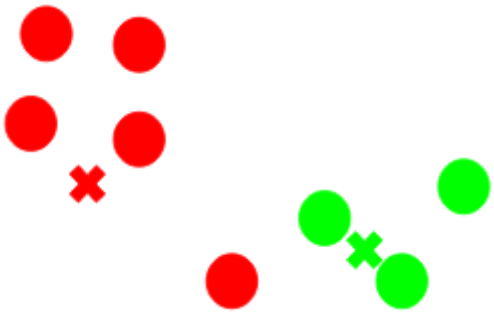
Once we have initialized the centroids, we assign each point to the closest cluster centroid:



Here we can see that the points which are closer to the red point are assigned to the red cluster whereas the points which are closer to the green point are assigned to the green cluster.

Step 4: Recompute the centroids of newly formed clusters

Now, once we have assigned all of the points to either cluster, the next step is to compute the centroids of newly formed clusters



Here, the red and green crosses are the new centroids.

Step 5: Repeat steps 3 and 4

We then repeat steps 3 and 4:

The step of computing the centroid and assigning all the points to the cluster based on their distance from the centroid is a single iteration. this is repeated till the centroid no longer changes, i.e. the solution converges.

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

Solution:

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

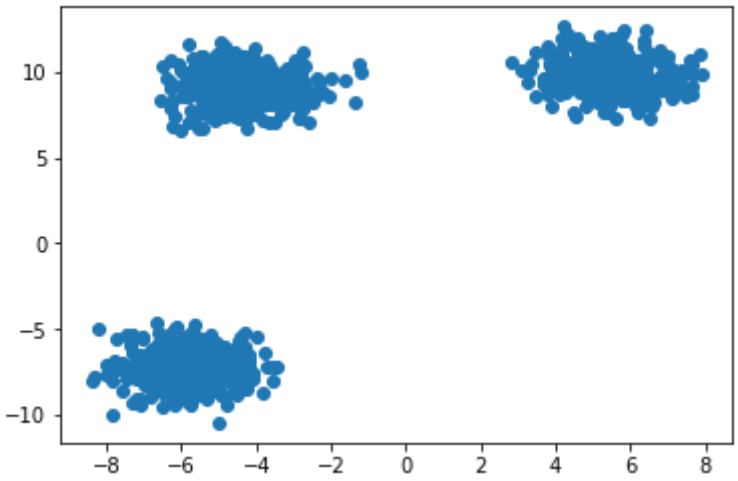
A cluster center is the representative of its cluster. The squared distance between each point and its cluster center is the required variation. The aim of k-means clustering is to find these k clusters and their centers while reducing the total error.

There are two methods to determine the optimal value of k as mentioned below

These methods are:

1. **The Elbow Method**
2. **The Silhouette Method**

Let us assume that we have a data set and a scatter plot is plotted for the values.



Clearly, the dataset has 3 clusters.

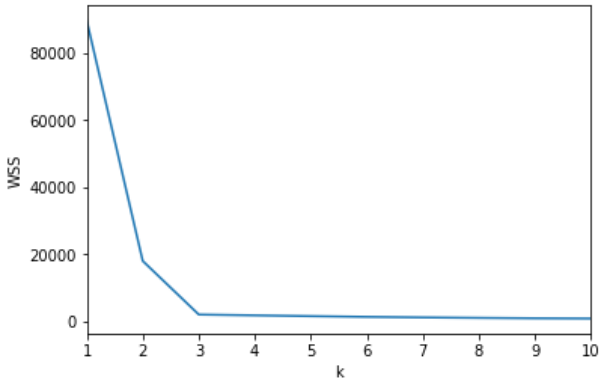
The Elbow Method

This is probably the most well-known method for determining the optimal number of clusters. 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.**

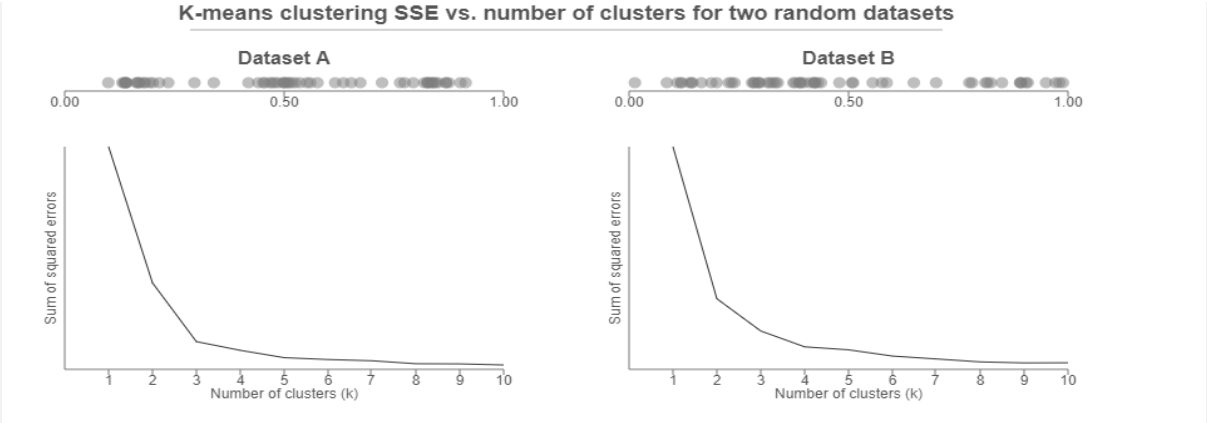
Steps involved are:

1. The Squared Error for each point is the square of the distance of the point from its representation i.e. its predicted cluster center.
2. The WSS score is the sum of these Squared Errors for all the points.
3. Any distance metric like the Euclidean Distance or the Manhattan Distance can be used.

On implementation we obtain the following plot for WSS-vs-k for our dataset.



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.



For Dataset A, the elbow is clear at k = 3. However, this choice is ambiguous for Dataset B. We could choose k to be either 3 or 4. 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

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

Note: s(i) is defined to be equal to zero if ***i*** is the only point in the cluster. This is to prevent the number of clusters from increasing significantly with many single-point clusters.

Here,

a(i) is the measure of similarity of the point ***i*** to its own cluster. It is measured as the average distance of ***i*** from other points in the cluster.

For each data point $i \in C_i$ (data point i in the cluster C_i), let

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} d(i, j)$$

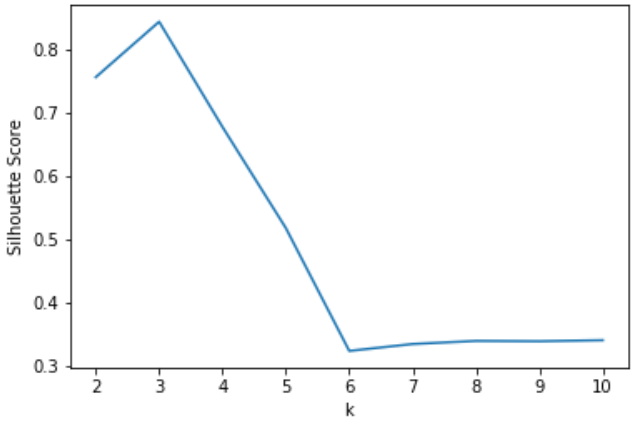
b(i) is the measure of dissimilarity of ***i*** from points in other clusters.

For each data point $i \in C_i$, we now define

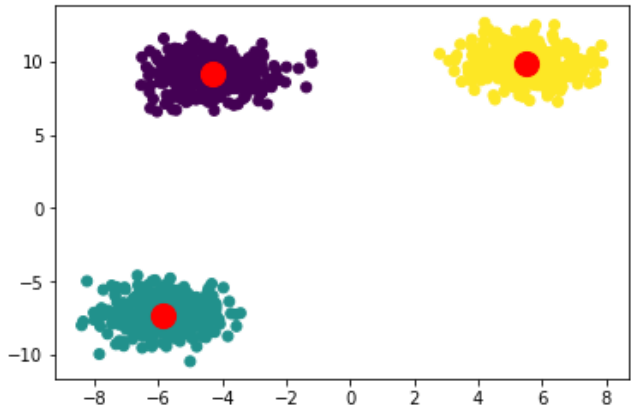
$$b(i) = \min_{i \neq j} \frac{1}{|C_j|} \sum_{j \in C_j} d(i, j)$$

d(i, j) is the distance between points ***i*** and ***j***. Generally, **Euclidean Distance** is used as the distance metric.

As mentioned earlier that a high Silhouette Score is desirable. The Silhouette Score reaches its ***global maximum at the optimal k***. This should ideally appear as a peak in the Silhouette Value-versus-k plot.
The plot for our dataset:



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



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.
Solution:
K-Means is an unsupervised learning algorithm which is widely used to cluster data into different groups-Means are distance-based algorithms. K-Means clusters the similar points together. The similarity here is defined by the distance between the points. Lesser the distance between the points, more is the similarity and vice versa.

All distance-based algorithms are affected by the scale of the variables. Let us consider data that has an age variable which tells about the age of a person in years and an income variable which tells the monthly income of the person in rupees:

ID	Age	Income (Rupees
1	25	80,000
2	30	100000
3	40	90000
4	30	50000
5	40	110000

Here the Age of the person ranges from 25 to 40 whereas the income variable ranges from 50,000 to 110,000. Let's try to find the similarity between observation 1 and 2. The most common way is to calculate the Euclidean distance and smaller this distance closer will be the points and hence they will be more similar to each other.

Euclidean distance is given by:

$$D = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_i - q_i)^2 + \cdots + (p_n - q_n)^2}.$$

Here,
n = number of variables
p1,p2,p3,… = features of first point
q1,q2,q3,… = features of second point

The Euclidean distance between observation 1 and 2 will be given as:

Euclidean Distance = [(100000–80000) ^2 + (30–25) ^2] ^ (1/2)

which will come out to be around 20000.000625.

It can be noted here that the high magnitude of income affected the distance between the two points. This will impact the performance of all distance-based model as it will give higher weightage to variables which have higher magnitude (income in this case). We do not want our algorithm to be affected by the magnitude of these variables. The algorithm should not be biased towards variables with higher magnitude. To overcome this problem, we can bring down all the variables to the same scale. One of the most common technique to do so is normalization where we calculate the mean and standard deviation of the variable. Then for each observation, we subtract the mean and then divide by the standard deviation of that variable:

$$z = \frac{x - \mu}{\sigma}$$

Apart from normalization, there are other methods too to bring down all the variables to the same scale. For example: Min-Max Scaling. Here the scaling is done using the following formula:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Let's see how normalization can bring down these variables to same scale and hence improve the performance of these distance-based algorithms. If we normalize the above data, it will look like:

ID	Age	Income (Rupees
1	-1.192	-0.260
2	-0.447	0.608
3	1.043	0.173
4	-0.447	-1.563
5	1.043	1.042

Let's again calculate the Euclidean distance between observation 1 and 2:

Euclidean Distance = [(0.608+0.260) ^2 + (-0.447+1.192) ^2] ^ (1/2)

This time the distance is around 1.1438.

We can clearly see that the distance is not biased towards the income variable. It is now giving similar weightage to both the variables. Hence, it is always advisable to bring all the features to the same scale for applying distance-based algorithms

Clustering models are distance-based algorithms, in order to measure similarities between observations and form clusters they use a distance metric. So, features with high ranges will have a bigger influence on the clustering. Therefore, standardization is required before building a clustering model.

e) Explain the different linkages used in Hierarchical Clustering.

Solution

The process of Hierarchical Clustering involves either clustering sub-clusters (data points in the first iteration) into larger clusters in a agglomerative (bottom-up) manner or dividing a larger cluster into smaller sub-clusters in a divisive (top-down manner).

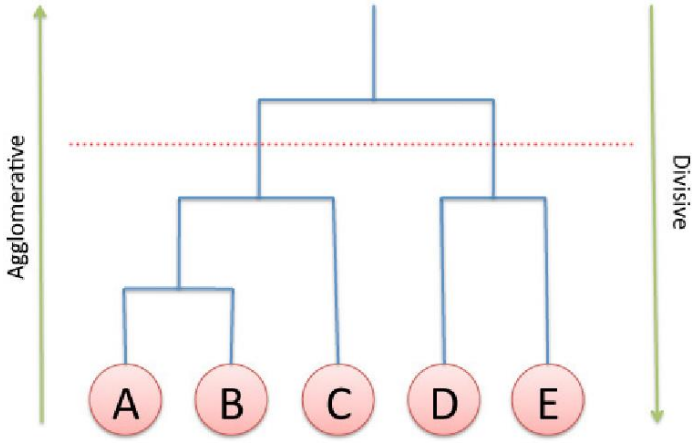
1. **Divisive**

Divisive hierarchical clustering works by starting with 1 cluster containing the entire data set. The observation with the highest average dissimilarity (farthest from the cluster by some metric) is reassigned to its own cluster. Any observations in the old cluster closer to the new cluster are assigned to the new cluster. This process repeats with the largest cluster until each observation is its own cluster.

2. **Agglomerative**

Agglomerative clustering starts with each observation as its own cluster. The two closest clusters are joined into one cluster. The next closest clusters are grouped together and this process continues until there is only one cluster containing the entire data set.

A hierarchical clustering is often represented as a dendrogram.



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: -

1. **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. These clusters can appear spread-out.

2. **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.

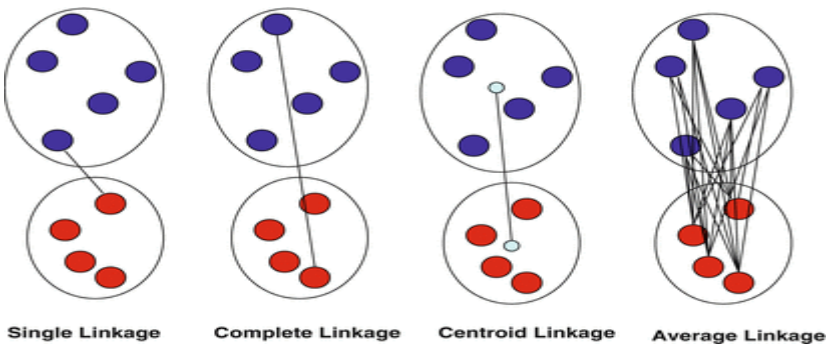
3. **Average-Linkage**

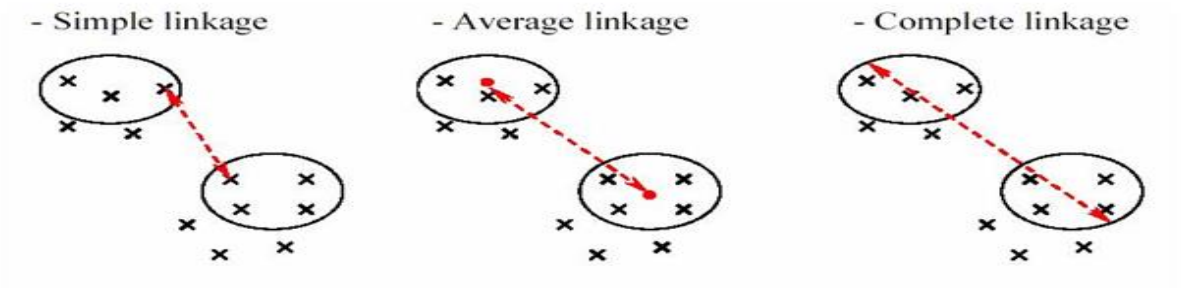
Average-linkage is where the distance between each pair of observations in each cluster are added up and divided by the number of pairs to get an average inter-cluster distance.

Average-linkage and complete-linkage are the two most popular distance metrics in hierarchical clustering.

4. **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.





Cluster distance measures

- Single link: $D(c_1,c_2) = \min_{x_1 \in c_1, x_2 \in c_2} D(x_1,x_2)$
  - distance between closest elements in clusters
  - produces long chains a→b→c→…→z
- Complete link: $D(c_1,c_2) = \max_{x_1 \in c_1, x_2 \in c_2} D(x_1,x_2)$
  - distance between farthest elements in clusters
  - forces "spherical" clusters with consistent "diameter"
- Average link: $D(c_1,c_2) = \frac{1}{|c_1|}\frac{1}{|c_2|}\sum_{x_1 \in c_1}\sum_{x_2 \in c_2} D(x_1,x_2)$
  - average of all pairwise distances
  - less affected by outliers
- Centroids: $D(c_1,c_2) = D\left(\left(\frac{1}{|c_1|}\sum_{x \in c_1}\vec{x}\right),\left(\frac{1}{|c_2|}\sum_{x \in c_2}\vec{x}\right)\right)$
  - distance between centroids (means) of two clusters
- Ward's method: $TD_{c_1 \cup c_2} = \sum_{x \in c_1 \cup c_2} D(x,\mu_{c_1 \cup c_2})^2$
  - consider joining two clusters, how does it change the total distance (TD) from centroids?