

Submitted by: Kanika Khattar Ahuja

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(why you took that many numbers of principal components, which type of Clustering produced a better result and so on)

Problem Statement

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. After the recent funding programmes, they have been able to raise around \$ 10 million. Now the CEO of the NGO needs to decide how to use this money strategically and effectively. The significant issues that come while making this decision are mostly related to choosing the countries that are in the direst need of aid.

You need to categorise the countries using some socio-economic and health factors that determine the overall development of the country. Then you need to suggest the countries which the CEO needs to focus on the most.

Solution: The following steps were taken to arrive at the final countries.

Step 1: Read and understand the data:

Step 2: Clean and visualize the data:

- No missing values were noticed.
- We do saw some outliers. As the data points are less so we did not be remove any outliers. The solution demands the name of countries that need financial aid, and deleting outliers might delete the country which is in need. Therefore, we did soft capping on the data.
- All points lying below 0.01 percentile will be assigned the value of 0.01 percentile
- All the points lying above 0.99 percentile will be assigned the value of 0.99 percentile

Step 3: Prepare the data for modeling

Step 3.1: Standardizing data: The data was scaled using standardScaler.

Step 3.2: Perform PCA and select the number of components:

- The need of PCA is because of the following reasons:
 - Initially all the variables are equally contributing towards explaining the variance. If we drop any feature, we will be losing information. Therefore,

we will be using PCA to reduce the dimensions before applying any model.

- We need to apply KMeans model and for any distance based algorithm there is a drawback. As we increase the number of dimensions the data points start looking equidistant from the cluster center. In such a case, K-Means clustering will fail to assign the data to its nearest cluster. Therefore, we will be using PCA to reduce the dimensions.
- A scree plot was formed to calculate the number of components. 95% of the variance was explained by 4 Principal Components and therefore we decided to create 4 Principal Components.

Step 3.3: Performing PCA with selected components:

- PCA was performed with 4 Principal Components. The maximum variance of 53.1% was explained by just PC1. Also the Components formed were not correlated with each other.

Step 4: Modelling with kMeans clustering

Step 4.1 Checking data compatibility with KMeans:

- Hopkins statistic was used to check the compatibility.
- On multiple runs the score was between 76 to 85
- The data was compatible for performing KMeans clustering

Step 4.2 Find out what should be the optimal number of clusters using SSD and silhouette score:

- In the plot for the elbow curve we saw that 3 or 4 clusters should be enough to create clusters of countries.
- Silhouette score was calculated and for For $n_clusters=4$, the silhouette score 0.35703 was the max
- Based on elbow curve and Silhouette score, $k=4$ was chosen.

Step 4.3 Performing KMeans with chosen number of clusters

Step 5: Cluster Profiling

Step 5.1 Analysis based on the cluster ids

- The 4 clusters are clearly visible and are separated mainly on the basis of PC1 as PC1 captures the maximum variance in the data.
- There were just 3 countries in cluster 3 and that seems to have very high income and very high gdpp with very low mortality rate
- Countries in cluster 2 have high gdpp, high income and low mortality rate.
- Countries in cluster 0 have low income and low gdpp and a comparatively higher child mortality rate than countries in cluster 2 and 3

- Countries in cluster 1 have a very low income and very low gdpp. The child mortality rate seems to be very high.
- Looking at the above analysis, we see that countries in cluster 1 are in need of financial aid,

Step 5.2 Analysis of a particular cluster:

- The countries were sorted based on the socio-economic factors. The sort order is selected based on the importance of the factor.

Step 6: Modeling with Hierarchical Clustering

- The dendrograms for linkage type simple, complete and average were created. Complete linkage gave a good dendrogram and was further used to divide the clusters.
- From the dendrogram, 5 clusters were clearly visible, therefore we created labels with 5 clusters.
- Clusters formed were somewhat similar with KMeans but there is a slight overlap between the clusters

Final Analysis:

Though the clusters are formed differently with KMeans and Hierarchical Clustering, the top 9 countries which seems to be in direst need of aid are the same. The CEO needs to focus on the following 9 countries

1. Liberia
2. Burundi
3. Congo, Dem. Rep.
4. Niger
5. Sierra Leone
6. Madagascar
7. Mozambique
8. Central African Republic
9. Malawi

Question 2: Clustering

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

Answer:

- Both KMeans clustering and hierarchical clustering is an unsupervised learning approach. It is used to partition data set into K distinct non overlapping clusters.
- Both K-Means and Hierarchical clustering are distance based algorithms which divide the data points into clusters.

- In K-Means, we need to decide on the value of K before performing the clustering but in hierarchical clustering, we need not decide on the value of k prior to running the algorithm. We can find the appropriate clusters by interpreting the dendrogram in hierarchical clustering.
- Hierarchical clustering is more computationally intensive but generally produces better clusters than K-Means.

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

Answer:

KMeans clustering is an unsupervised learning approach. It is used to partition data set into K distinct non overlapping clusters. Given the value of K, the K-Means clustering algorithm assigns each data point to exactly one of the K clusters.

The steps of the algorithm to create K clusters are as follows:

1. **Initialisation:** Randomly K points are chosen which acts as initial cluster centroids. The points can be randomly chosen from any of the data points from observations or it can be a totally different point.
2. **Assignment:** Each data point is then assigned to a particular cluster based on the closest squared Euclidean distance with the cluster centroids.

$$C_k = \text{Argmin} \left\{ \sum_{k=1}^K \sum_{i=1}^N (x_i - \mu_k)^2 \right\}$$

where

C_k = set containing the observations in k^{th} cluster

K = total number of desired clusters

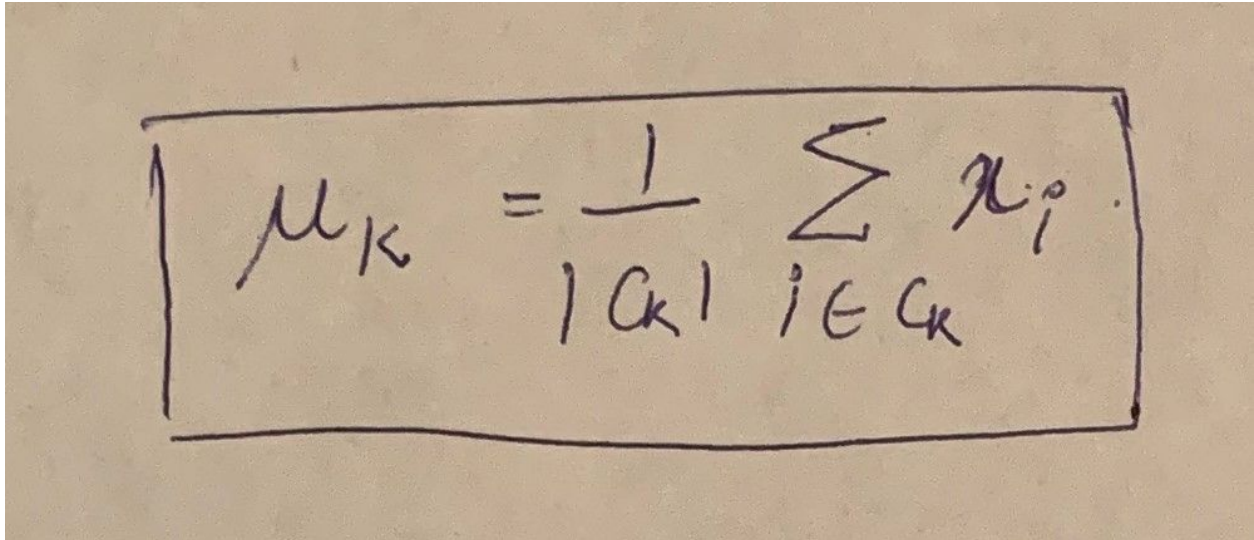
N = total number of observations

The data points are assigned such that:

Each data point belongs to atleast one of the K clusters

No data point belongs to more than one cluster.

3. **Optimisation:** After the data points are assigned to the clusters, the cluster center is computed again for each of the K clusters by taking the mean of distance of cluster centroid with its assigned members. Once the new centroids are calculated, again the assignment step is done and members are allotted to the clusters.



A handwritten formula for calculating the new centroid of a cluster, enclosed in a hand-drawn rectangular box. The formula is:
$$\mu_k = \frac{1}{|C_k|} \sum_{i \in C_k} x_i$$

4. **Iteration:** The above two steps of Assignment and Optimisation are repeated until there is no change in the clusters or possibly until the algorithm converges.

The following figure explains the above steps of the algorithm

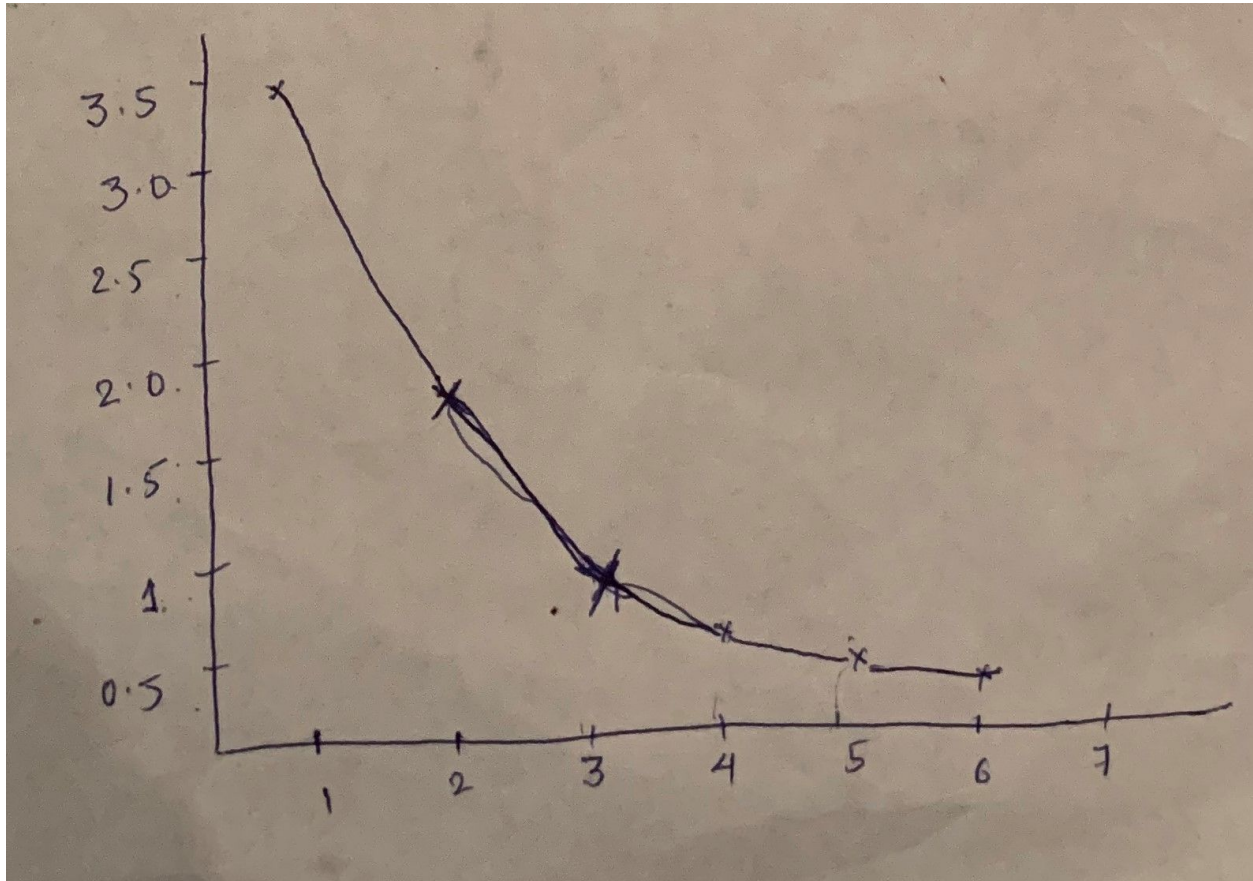
c) How is the value of 'k' chosen in K-means clustering? Explain both the statistical as well as the business aspect of it.

Answer:

There are 2 methods that help in choosing the value of K for K-Means clustering.

Method 1: Elbow Method:

- It helps in deciding the value of K by interpreting and validating the consistency within the cluster.
- It runs the K-Means clustering on the dataset for a range of values of K. Along with this, it calculates the sum of squared Errors for each value of K.
- The line chart of SSE for each value of K is plotted. An arm like plot is formed as shown in the figures and the elbow of the arm is the value of K that is supposed to give best clusters.
- The goal is to choose a small value of K which still has a low SSE.



If the data is not very well clustered, Elbow method might not work well. We have another method that might be considered.

Method 2: Silhouette Analysis

- It runs the K-Means clustering on the dataset for a range of values of K. Along with this, it calculates a Silhouette Value which is a measure of how similar a data point is to its own cluster (cohesion) compared to other clusters(separation)
The Silhouette Value is given by

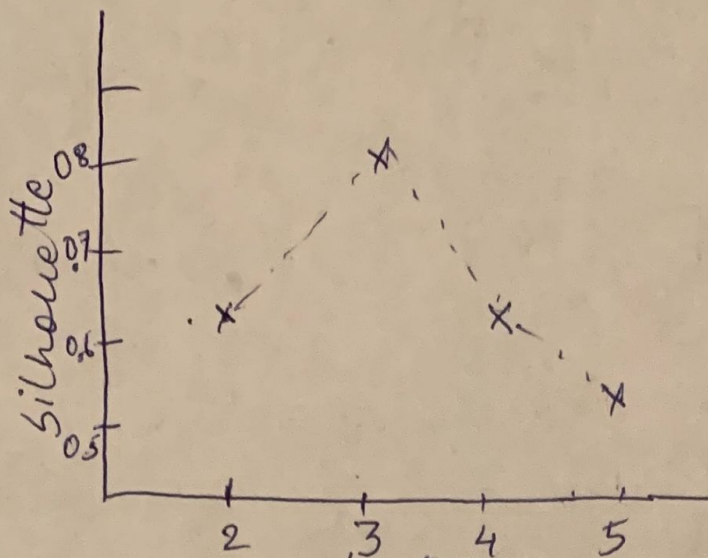
$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

where

a_i is the avg distance from own clusters

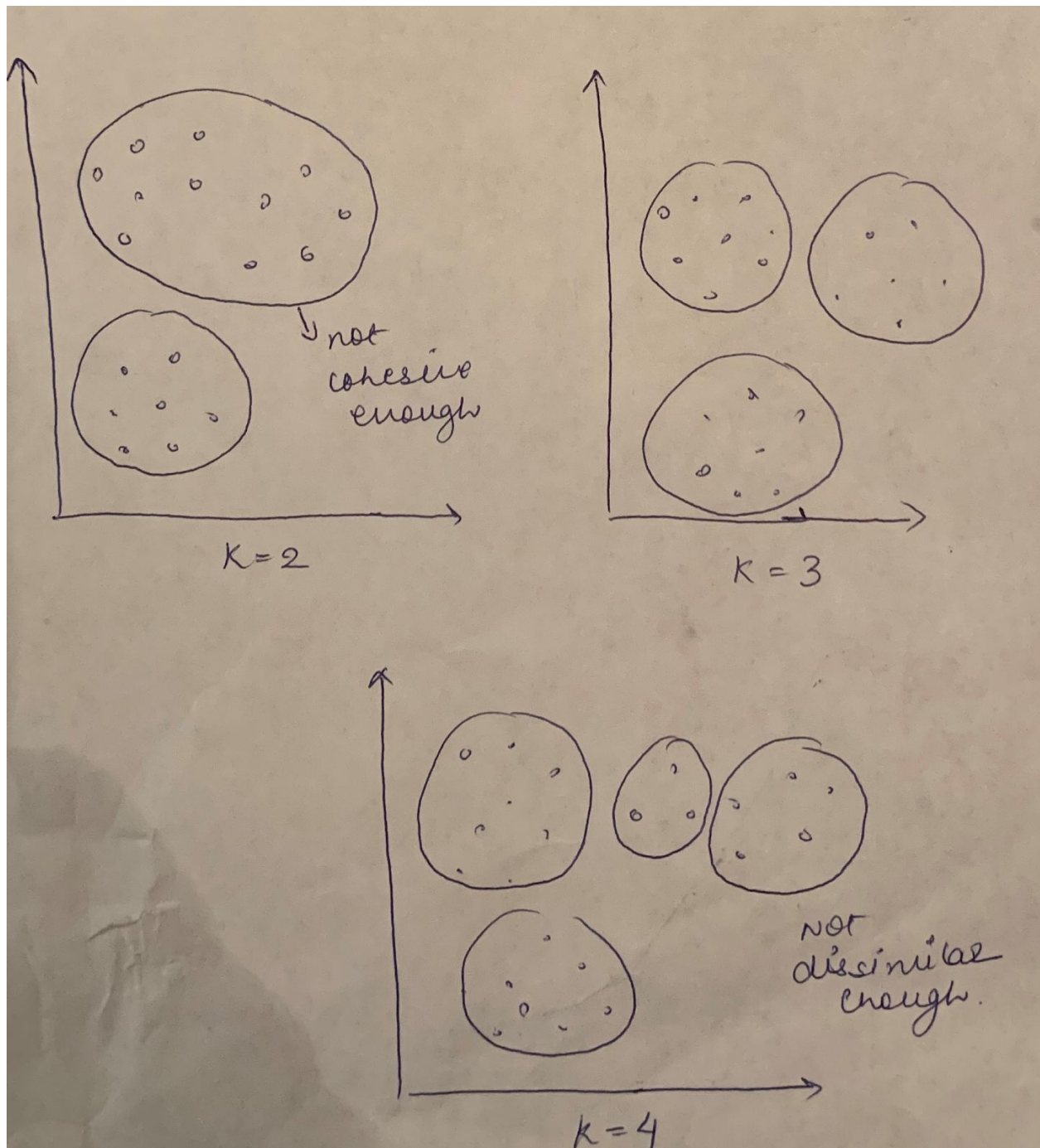
b_i is the avg distance from nearest neighbour clusters.

- A curve of silhouette value for each value of K is plotted.
- The location of the maximum is considered as the appropriate number of clusters.



$K = 3$ gives the max silhouette.

Choosing the value of K is of utmost importance as lower values of K can cause clusters not to be cohesive enough and higher values of K can lead to clusters not dissimilar enough.



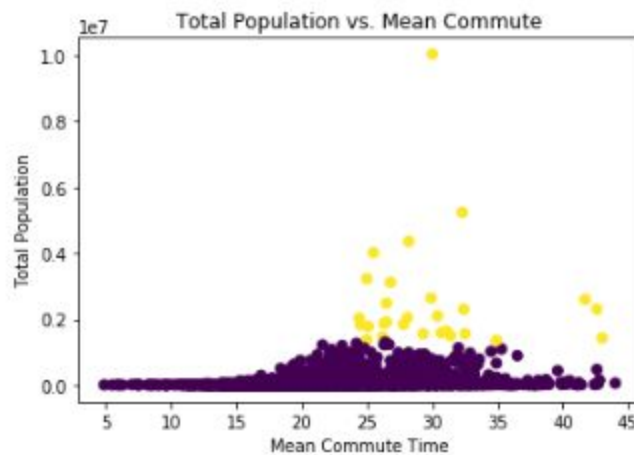
d) Explain the necessity for scaling/standardisation before performing Clustering.

Answer:

Clustering algorithm uses the Euclidean distance metric to cluster the data points. Therefore, if the data is not scaled, attributes with larger range of values might out-weigh the attributes with smaller range. Thus there is a need of scaling down of all attributes to the same normal scale. Scaling helps in making the attributes **unit-free and uniform**

Consider an example of clustering the data based on total population and mean commute time to split the countries into 2 groups. The units of these attributes are very different.

If we perform clustering without standardising the data, total population will be the primary driver for dividing the countries in two groups.



After standardization both the attributes - total population and Mean Commute time seem to have an influence on cluster formation

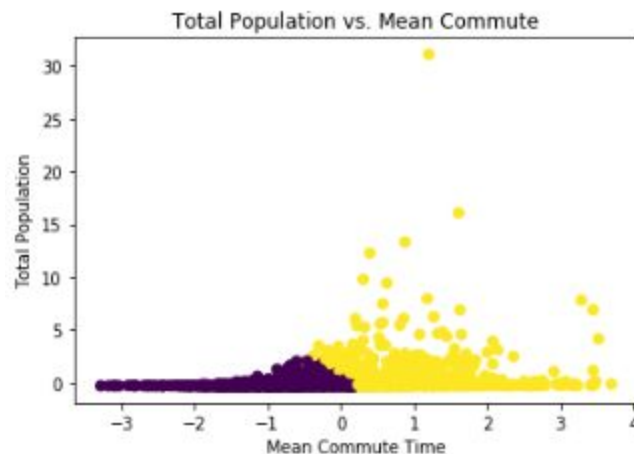


Image source:

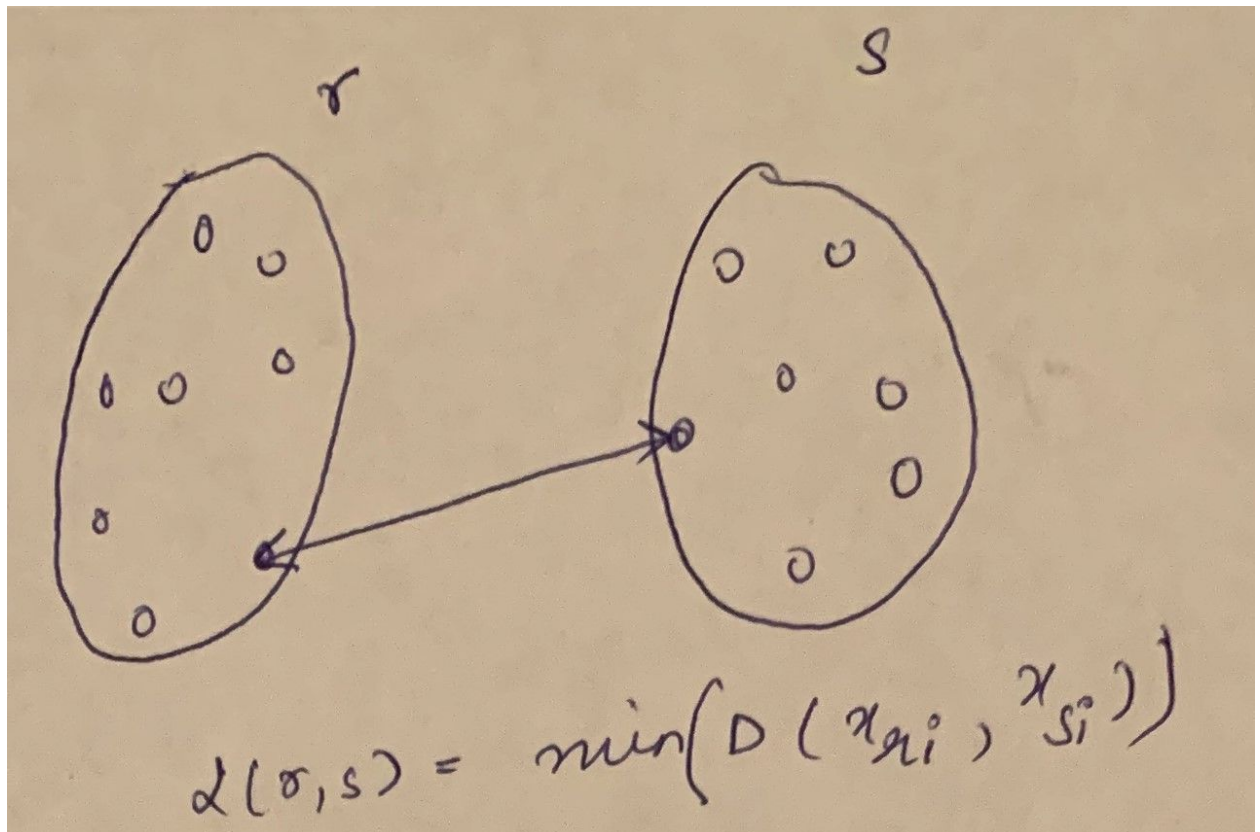
<https://community.alteryx.com/t5/Alteryx-Designer-Knowledge-Base/Standardization-in-Cluster-Analysis/ta-p/302296>

e) Explain the different linkages used in Hierarchical Clustering.

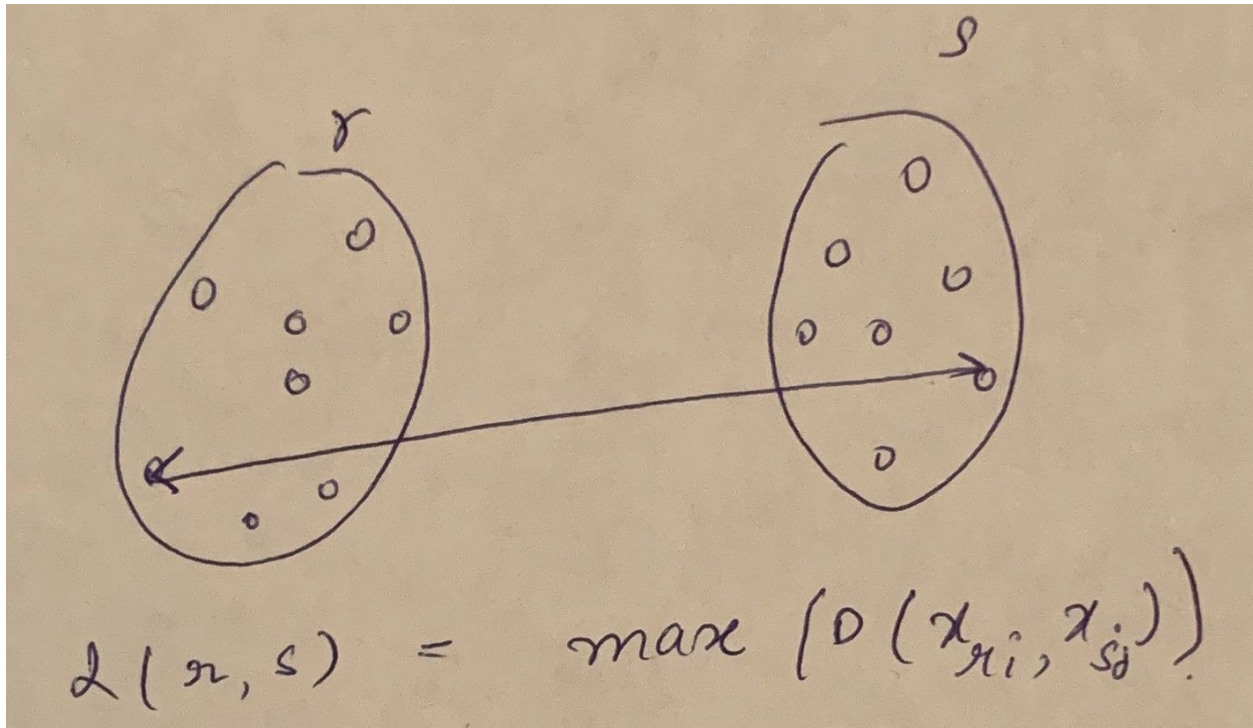
Answer:

In all the clustering algorithms it is important to determine the distance between each cluster using a distance function. These linkages between two clusters can be measured by following distance functions based on which different linkages are defined in Hierarchical clustering.

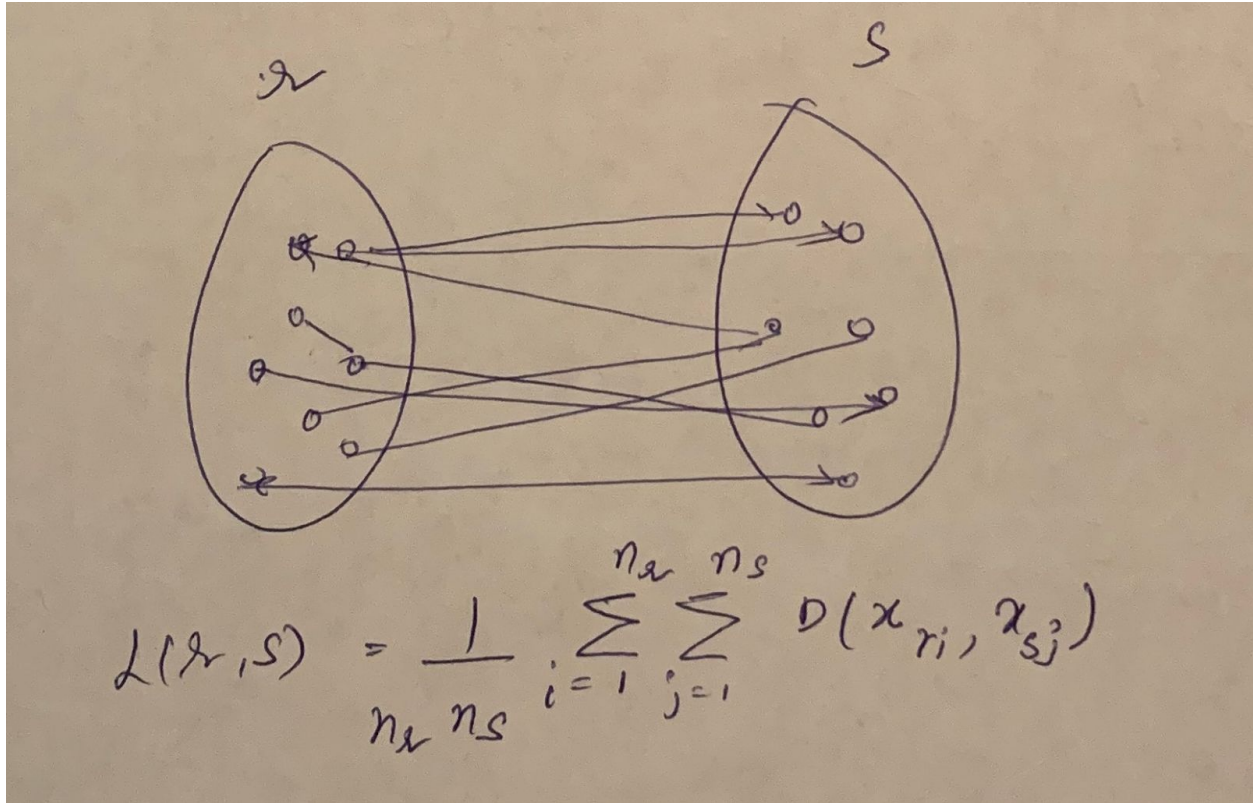
1. **Single Linkage:** The distance between two clusters is defined by the shortest distance between two points in each cluster. Therefore, in order to merge two groups only one pair of points is required to be near to each other. The clusters can be too spread out and not compact enough.



2. **Complete linkage:** The distance between two clusters is defined as the longest distance between two points in each cluster. It produces clusters that are compact but not far enough apart.



3. **Average Linkage:** The distance between two clusters is defined as the average distance between each point in one cluster to every other point in the other cluster. Clusters tend to be relatively compact as well as relatively apart.

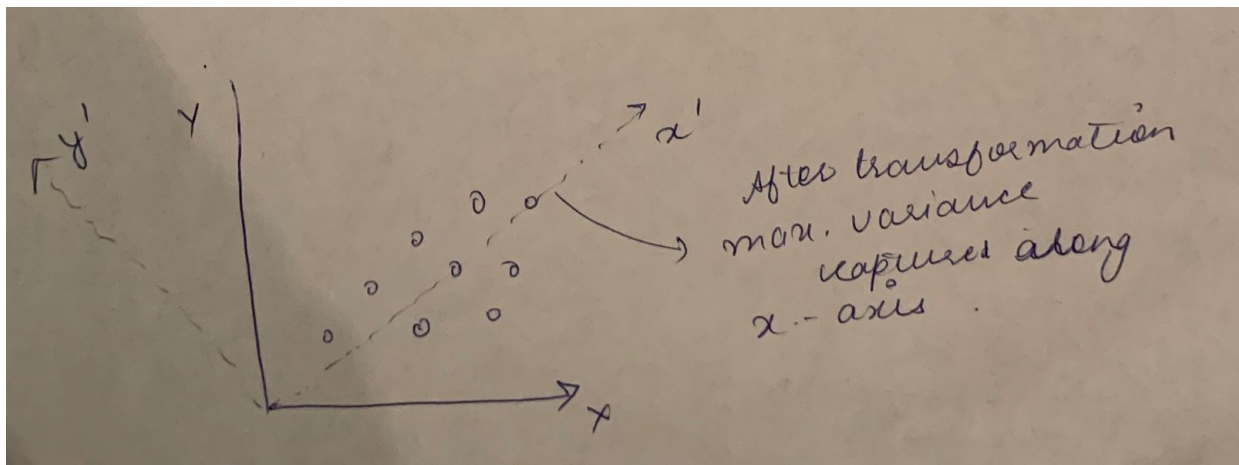


Question 3: Principal Component Analysis

a) Give at least three applications of using PCA

1. **Dimensionality Reduction:** One of the major application of PCA is dimensionality reduction. Having large number of dimensions in modelling increases complexity and also need more processing power. PCA helps in finding the hidden dimension in the data. As a result it provides the Principal Components which are the linear combination of original features and are uncorrelated to each other.

In the following figure, consider the data with 2 Dimensions x and y . The PCA will form the new axis also defined by basis and will transform the data onto different axis (x', y') such that the data points can only be explained using x' .



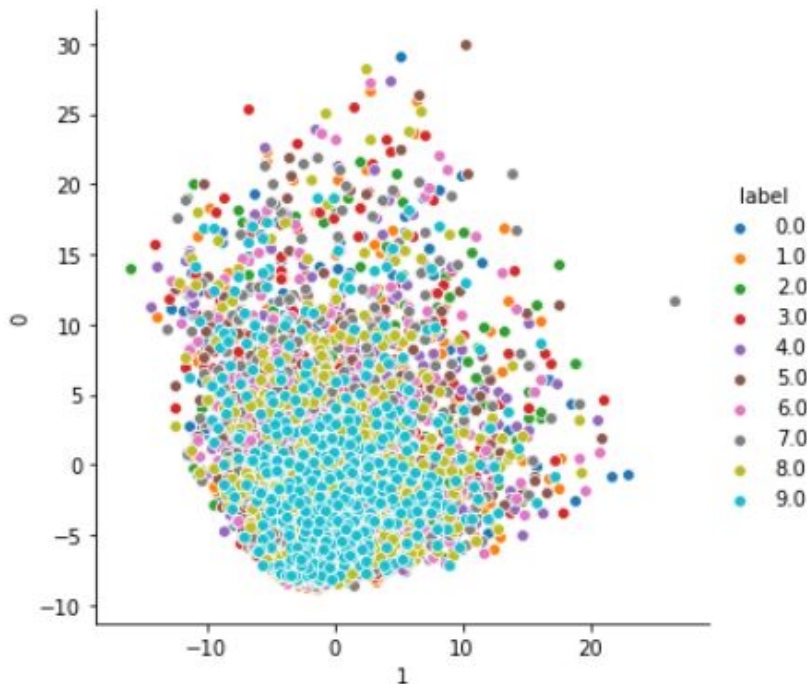
2. Data Visualisation

As discussed above, PCA helps in reducing the dimensions. Once the dimensions are reduced, it becomes easier to visualize the data.

For example, Considering the MNIST data for digit recognition, the original data consists of 785 columns where each column represent the value of a pixel in the digit image.

Visualizing the entire data with these many columns is very difficult. With the help of PCA we can reduce the data into 2 dimensions.

By looking at the following figure which is the scatter plot of two Principal components created on the MNSIT data, we can easily visualize how similar the digits 0 and 9.



The overlap of 0 and 9 labeled points is the maximum.

3. **Image compression:** Once we reduce the dimension of the data, the original images when transformed into lesser dimensions is compressed.

Given an image with original matrix of pixels represented by RGB color values, PCA reduces the dimensions of the matrix, to reform the image that retains its qualities but is smaller in weight.

The following image shows how the quality of the image is affected upon compressing using PCA with 50 components and 37 components.



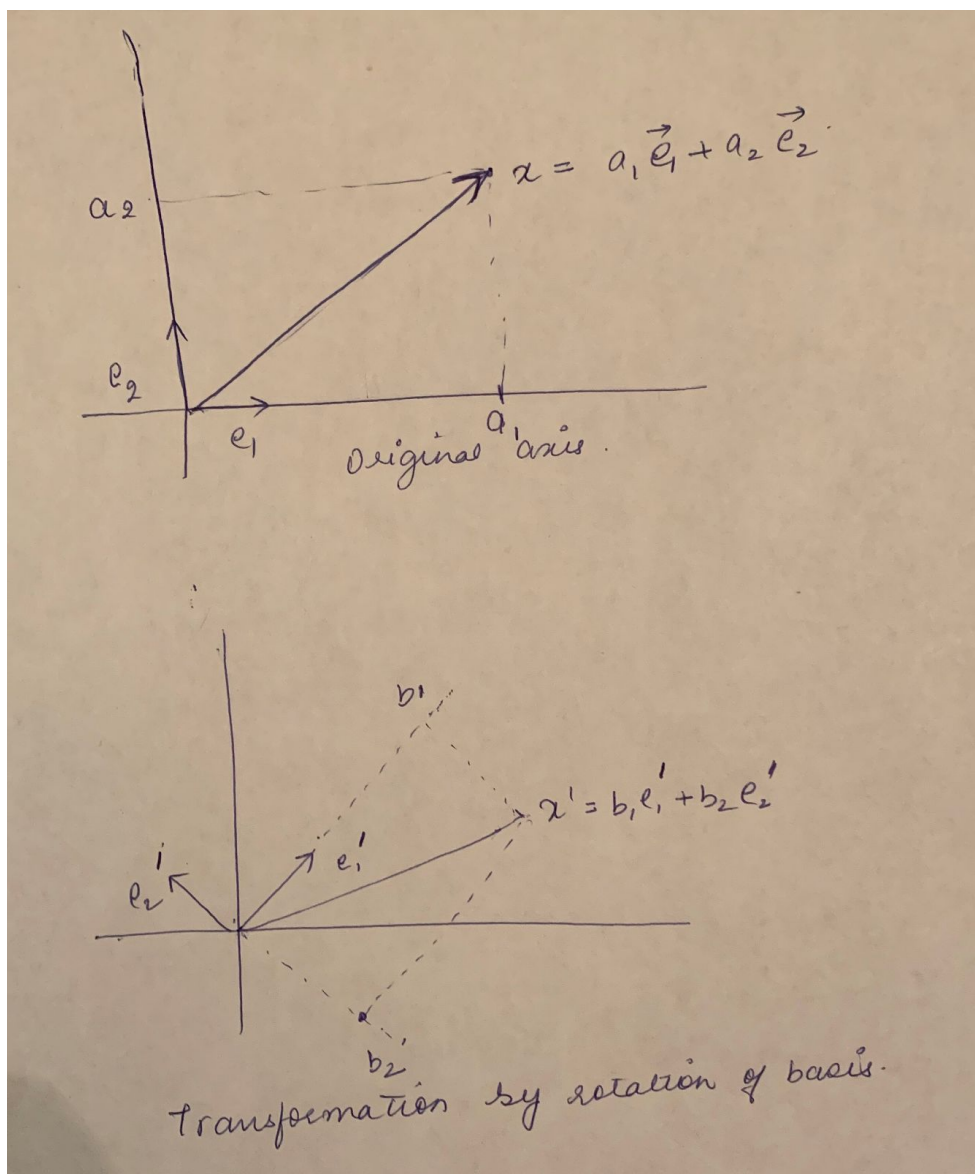
Image source: http://www.math.mcgill.ca/yyang/regression/extra/PCA_Demo

b) Briefly discuss the 2 important building blocks of PCA - Basis transformation and variance as information.

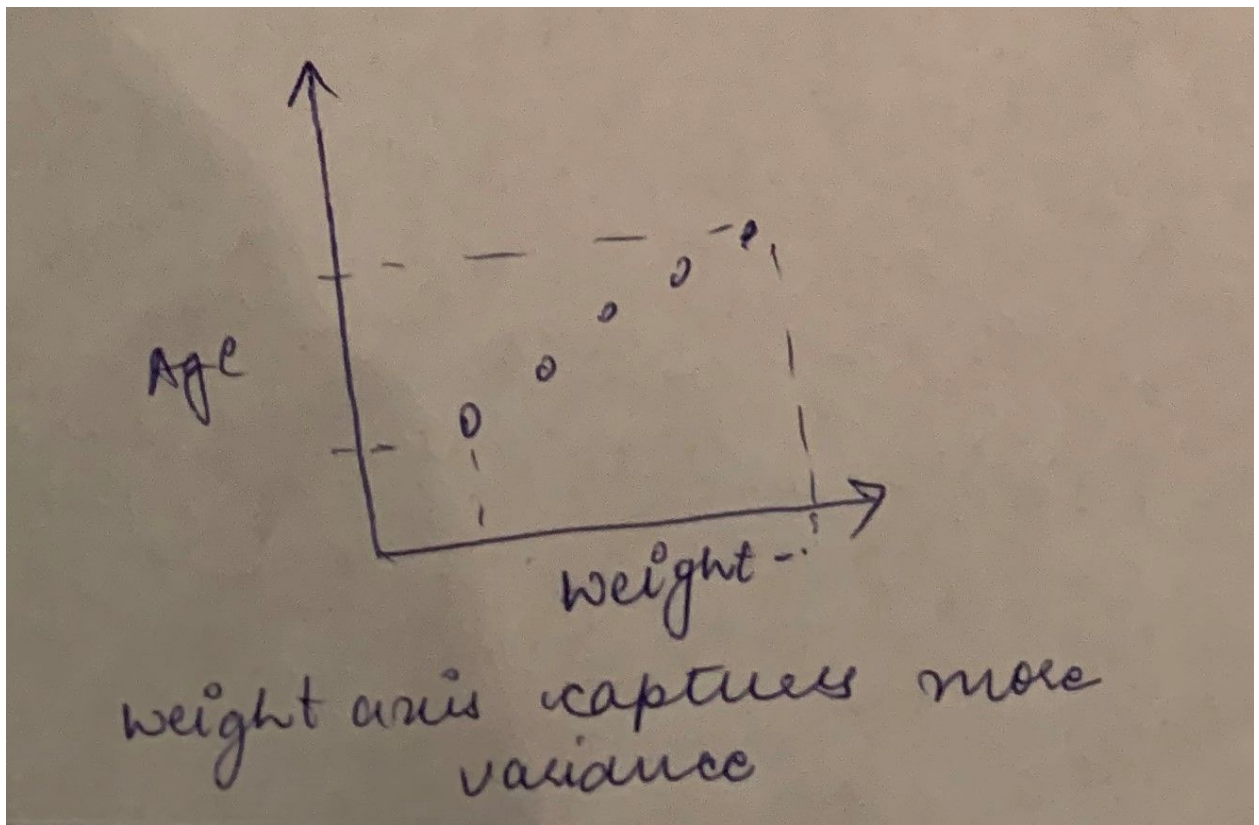
1. PCA - Basis Transformation

A basis vector space of dimension n is a set of n vectors ($a_1, a_2, a_3 \dots a_n$) such that every vector in the space can be represented as a unique linear combination of basis vectors. Coordinate wise representation of vectors and operators taken with respect to one basis can be easily transformed to their equivalent representations with respect to another basis.

This process of converting the information from one set of basis to another is called the basis transformation.



2. **Variance as information:** The variable that captures variance in the data are the variables that capture information in the data. Thus how important a column is checked by its variance values



PCA changes the basis vector in such a way that the new basis vector capture the maximum variance or information.

c) State at least three shortcomings of using Principal Component Analysis.

1. **The independent variables becomes less interpretable:** PCA transforms the original features into Principal Components which are the linear combination of the original features. The Principal Components are not as readable and interpretable as original features.
2. **Information Loss:** Principal Components tries to cover maximum variance among the features on a dataset, But if we don't select the number of components with care, it may miss some information as compared to the original list of features.
3. **Data Standardization is a must before PCA:** Data Standardization is a must before implementing PCA. If not done, PCA will not be able to find the optimal Principal Components. For example if a feature set has data expressed in units of Kilograms, Light years are millions, the variance scale is huge in the data set. If PCA is applied, the resultant Principal Components will be biased towards features with high variance.