ASSIGNMENT

- **ASSIGNMENT CLUSTERING:**

1. 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)

- First read the file and clean the data by converting the percentage column into absolute values. Then applying PCA to the data frame to reduce the variables of the data frame.
  Before PCA we scaled the data so that values in all columns comes to same scale.
  After performing PCA we get 4 components.
  After PCA we perform outliers and remove rows by using Inter quartile range(IQR).
  Hopkins value come as 0.7 and shows that data is good for clustering.
  Now, K-Means Clustering process is used for modelling.
  Through Silhouette Analysis and elbow tree we get K = 4
  And K = 5. First use K = 4 and K = 5 and then on comparison it is seen that there is not much difference.
  Through K-Means it comes to know that clusters 1 and 3 are most affected.
  After performing K-Means analysis we did Hierarchical Analysis on the data through that data we come to now that clusters 1 and 4 are mostly affected.
  At the end we concluded that through both the methods same countries are coming which are needed to be looked after by the HELP Organisation.

- **CLUSTERING**

1. **Compare and contrast K-Means and Hierarchical clustering.**

K-Means Clustering - It uses centroids, K- differently randomly initiated points in the data, and assigns every data point to the nearest centroid. After every point has been assigned, the centroid is moved to the average of all the points assigned to it. Then process repeats itself. The algorithm is done when no point changes assigned centroid.

Hierarchical Clustering - 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 after the fact of 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.

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

- Initialize K random centroids.

    - You could pick K random data points and make those your starting points.
    - Otherwise, you pick K random values for each variable.
- For every data point, look at which centroid is nearest to it.
    - Using some sort of measurement like Euclidean or Cosine distance.
- Assign the data point to the nearest centroid.
- For every centroid, move the centroid to the average of the points assigned to that centroid.
- Repeat the last three steps until the centroid assignment no longer changes.
    - The algorithm is said to have "converged" once there are no more changes.

These centroids act as the average representation of the points that are assigned to it. This gives you a story almost right away. You can compare the centroid values and tell if one cluster favours a group of variables or if the clusters have logical groupings of key variables.

3. How is the value of K is chosen in K- means clustering? Explain both and statistical as well as business aspect of it.

There is a method known as **elbow method** which is used to determine the optimal value of K to perform the K-Means Clustering Algorithm. The basic idea behind this method is that it plots the various values of cost with changing *k*. As the value of **K** increases, there will be fewer elements in the cluster. So average distortion will decrease. The lesser number of elements means closer to the centroid. So, the point where this distortion declines the most is the elbow point.

Business Aspect:

The K-means clustering algorithm is used to find groups which have not been explicitly labelled in the data. This can be used to confirm business assumptions about what types of groups exist or to identify unknown groups in complex data sets. Once the algorithm has been run and the groups are defined, any new data can be easily assigned to the correct group.
This is a versatile algorithm that can be used for any type of grouping. Some examples of use cases are:

- Behavioural segmentation:
  - Segment by purchase history
  - Segment by activities on application, website, or platform
  - Define personas based on interests
  - Create profiles based on activity monitoring
- Inventory categorization:
  - Group inventory by sales activity
  - Group inventory by manufacturing metrics
- Sorting sensor measurements:
  - Detect activity types in motion sensors
  - Group images
  - Separate audio
  - Identify groups in health monitoring
- Detecting bots or anomalies:
  - Separate valid activity groups from bots
  - Group valid activity to clean up outlier detection

In addition, monitoring if a tracked data point switches between groups over time can be used to detect meaningful changes in the data.

4. **Explain the necessity for scaling/standardisation before performing Clustering.**

n statistics, standardization (sometimes 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.

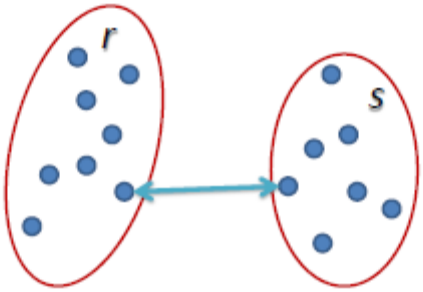
Often performed as a pre-processing step, particularly for cluster analysis, standardization may be important if you are working with data where each variable has a different unit (e.g., inches, meters, tons and kilograms), or where the scales of each of your variables are very different from one another (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.

When you are working with data where each variable means something different, (e.g., age and weight) the fields are not directly comparable. One year is not equivalent to one pound and may or may not have the same level of importance in sorting a group of records. In a situation where one field has a much greater range of value than another (because the field with the wider range of values likely has greater distances between values), it may end up being the primary driver of what defines clusters. Standardization helps to make the relative weight of each variable equal by converting each variable to a unitless measure or relative distance.

5. Explain the different linkages used in Hierarchical Clustering?

Single Linkage

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



$$L(r,s) = \min(D(x_{ri}, x_{sj}))$$

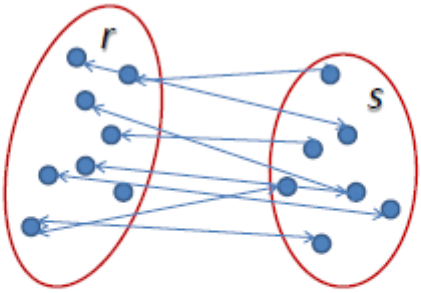
Complete Linkage

In complete linkage hierarchical clustering, the distance between two clusters is d
the *longest* distance between two points in each cluster. For example, the distance
"s" to the left is equal to the length of the arrow between their two furthest points.

$$L(r,s) = \max(D(x_{ri}, x_{sj}))$$

Average Linkage

In average linkage hierarchical clustering, the distance between two clusters is de
the average distance between each point in one cluster to every point in the other
For example, the distance between clusters "r" and "s" to the left is equal to the a
length each arrow between connecting the points of one cluster to the other.



$$L(r,s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} D(x_{ri}, x_{sj})$$

- **PRINCIPAL COMPONENT ANALYSIS**

1. **Give at least three applications of using PCA?**

a. The primary application of PCA is dimension reduction. If you have high Dimensional data, PCA allows you to reduce the dimensionality of the data so, most of the variation that exists in your data across many high dins captured in few dimensions.

b. PCA aims to orthogonally transform correlated variables to a smaller set of uncorrelated variables (principal components). Hence, this will the **multicollinearity.**

c. It is also used for finding patterns in data of high dimension in the field of data mining, bioinformatics, psychology, etc.

d. Normalize the data

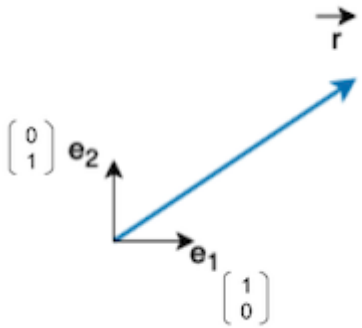
PCA is used to identify the components with the maximum variance, and contribution of each variable to a component is based on its magnitude variance. It is best practice to normalize the data before conducting a PC unscaled data with different measurement units can distort the relative comparison of variance across features.

2. **Briefly discuss the 2 important building blocks of PCA - Basis transform and variance as information?**

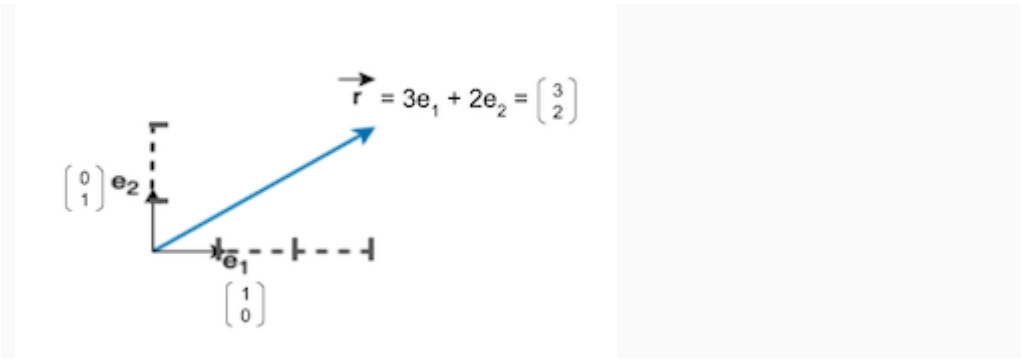
- **Basis Transformation:**

Basis change

Consider a vector ***r*** in a 2-dimensional space R^2. The space R^2 can be defin an arbitrary set of orthogonal, unit length vectors e1 and e2.



Now, looking at vector ***r*** with respect to coordinates e1 and e2, it can be represented as follows,



We initially denoted vectors e1 and e2 as an 'arbitrary' set of vectors. By convention, e1 and e2 form a standard coordinate set as they are of unit length and orthogonal to each other. When vector ***r*** is described with respect to a set of vectors e or basis e, it can be denoted as ***r***e (read vector r with basis e). However, the same vector r can still be described with respect to another set of coordinates, example b1 and b2. In this new coordinate system, the numbers of vector rb (read vector r with basis b) will be different; represented as [x,y] in the figure below.

Vectors b1 and b2 in the figure are described with respect to standard coordinates or basis e. Note that vectors b1 and b2 do not have to be orthogonal to each other (in this example, they are orthogonal).

VARIANCE:

The total variance is the sum of variances of all individual principal components.

The fraction of variance explained by a principal component is the ratio between the variance of that principal component and the total variance.

For several principal components, add up their variances and divide by the total variance.

3. **State at least three shortcomings of using Principal Component Analysis?**

Independent variables become less interpretable: After implementing PCA on the dataset, your original features will turn into Principal Components. Principal Components are the linear combination of your original features. Principal Components are not as readable and interpretable as original features.

2. Data standardization is must before PCA: You must standardize your data before implementing PCA, otherwise PCA will not be able to find the optimal Principal Components.

For instance, if a feature set has data expressed in units of Kilograms, Light years, or Millions, the variance scale is huge in the training set. If PCA is applied on such a feature set, the resultant loadings for features with high variance will also be large. Hence, principal components will be biased towards features with high variance, leading to false results.

Also, for standardization, all the categorical features are required

to be converted into numerical features before PCA can be applied.

PCA is affected by scale, so you need to scale the features in your data before applying PCA. Use StandardScalar from Scikit Learn to standardize the dataset features onto unit scale (mean = 0 and standard deviation = 1) which is a requirement for the optimal performance of many Machine Learning algorithms.

3. Information Loss: Although Principal Components try to cover maximum variance among the features in a dataset, if we don't select the number of Principal Components with care, it may miss some information as compared to the original list of features.