Ques 1: Compare and contrast K-means Clustering and Hierarchical Clustering.

Ans 1:

In k-Means clustering, the desired number of clusters are predefined (this is the 'k' value).

k-means will often give unintuitive results if

1.Your data is not well-separated into sphere-like clusters

2.You pick a 'k' not well-suited to the shape of your data, i.e. you pick a value too high or too low

3.You have weird initial values for your cluster centroids (one strategy is to run a bunch of k-means algorithms with random starting centroids and take some common clustering result as the final result).

In contrast, hierarchical clustering has fewer assumptions about ethe distribution of your data –

a. the only requirement (which k-means also shares) is that a distance can be calculated each pair of data points.
b. Hierarchical clustering typically 'joins' nearby points into a cluster, and then successively adds nearby points to the nearest group.
c. You end up with a 'dendrogram', or a sort of connectivity plot.
d. You can use that plot to decide after the fact of how many clusters your data has, by cutting the dendrogram at different heights.
e. 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.
f. Hierarchical clustering can be more computationally expensive but usually produces more intuitive results.

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

Ans 2: The way k means algorithm works is as follows:

1. Specify number of clusters *K*.
2. Initialize centroids by first shuffling the dataset and then randomly selecting *K* data points for the centroids without replacement.

   Keep iterating until there is no change to the centroids. i.e. assignment of data points to clusters isn't changing. The algorithm's inner-loop iterates over two steps:

   a. Assign each observation Xi to the closest cluster centroid μk

   b. Update each centroid to the mean of the points assigned to it.

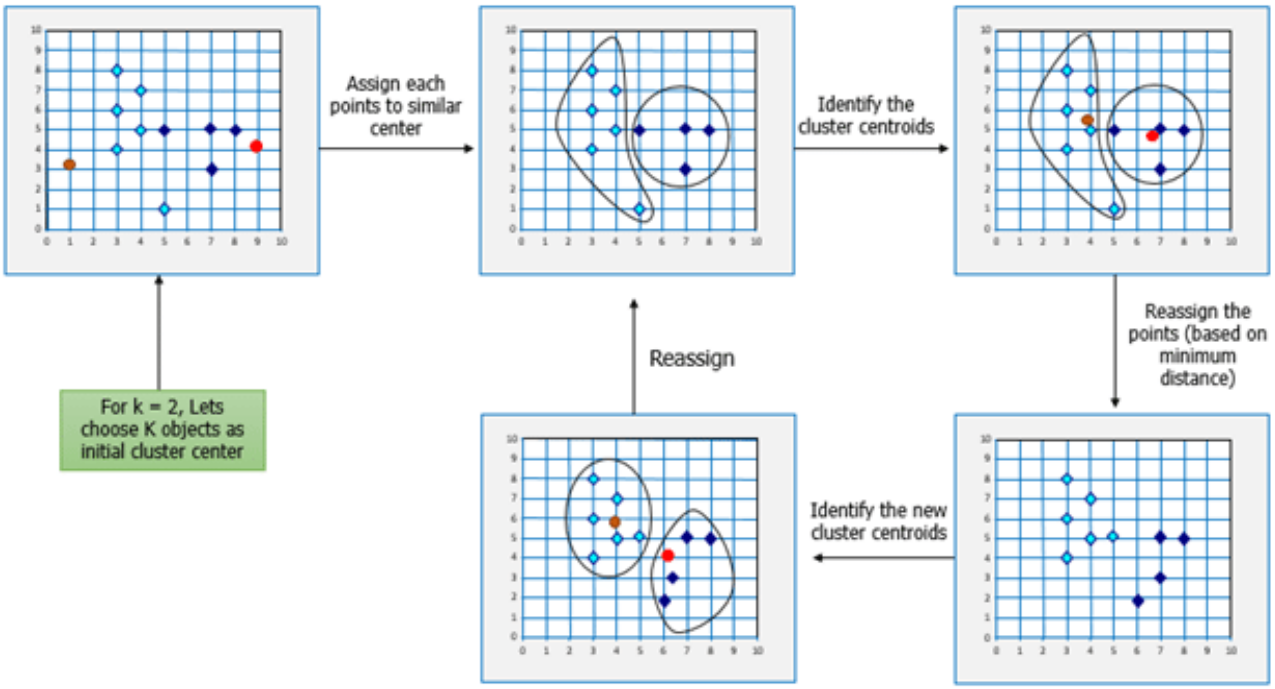
3. Compute the sum of the squared distance between data points and all centroids.
4. **Assign each data point to the closest cluster (centroid).:** In the assignment step, we assign every data point to K clusters.

   The algorithm goes through each of the data points and depending on which cluster is closer, in our case, whether the green cluster centroid or the blue cluster centroid; It assigns the data points to one of the 2 cluster centroids.

   Now having assigned each data point to a cluster, now we need to recompute the cluster centroids.

5. **Compute the centroids for the clusters by taking the average of the all data points that belong to each cluster.:** In the optimisation step, the algorithm calculates the average of all the points in a cluster and moves the centroid to that average location.

   The process of assignment and optimisation is repeated until there is no change in the clusters or possibly until the algorithm converges.


For k = 2, Lets choose K objects as initial cluster center
Assign each points to similar center
Identify the cluster centroids
Reassign the points (based on minimum distance)
Identify the new cluster centroids
Reassign

Q 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 c) The value of k in k means is chosen in following ways:

Statistical Aspect

Elbow method

It 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**.

Silhouette analysis

The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The silhouette ranges from −1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. If most objects have a high value, then the clustering configuration is appropriate. If many points have a low or negative value, then the clustering configuration may have too many or too few clusters.

Business Aspect

Business aspect is very important aspect and we should not only depend on the Statistical aspect to consider the number of clusters.

The number of clusters we get, should signify some meaning also. We cannot just depend on the statistical value. If we want to consider the number of cluster as per the business aspect and it is meaningful in some way then we should consider that particular number of clusters.

Q d) Explain the necessity for scaling / standardization before performing Clustering

Ans d

In 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.

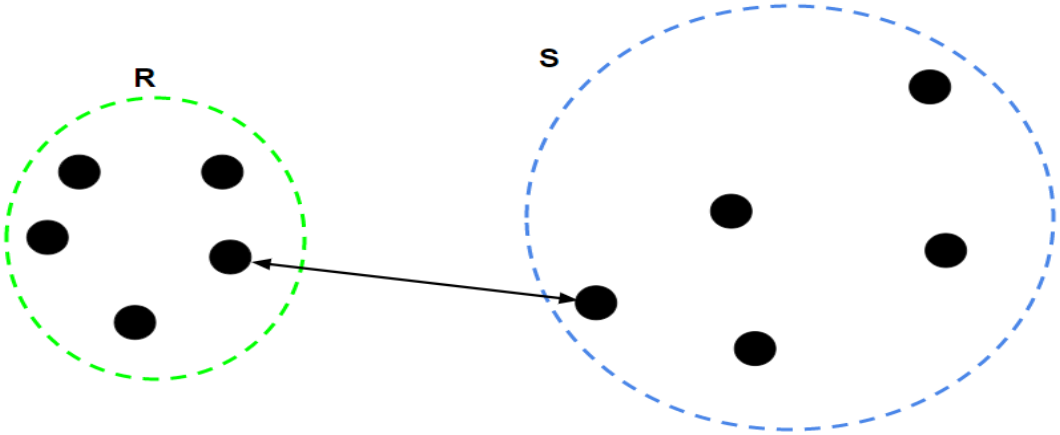
Standardization is an important step of Data preprocessing because:

a) 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).
b) It controls the variability of the dataset, it convert data into specific range using a linear transformation which generate good quality clusters and improve the accuracy of clustering algorithms.
c) The reason this importance is particularly high in cluster analysis is because groups are defined based on the distance between points in mathematical space.

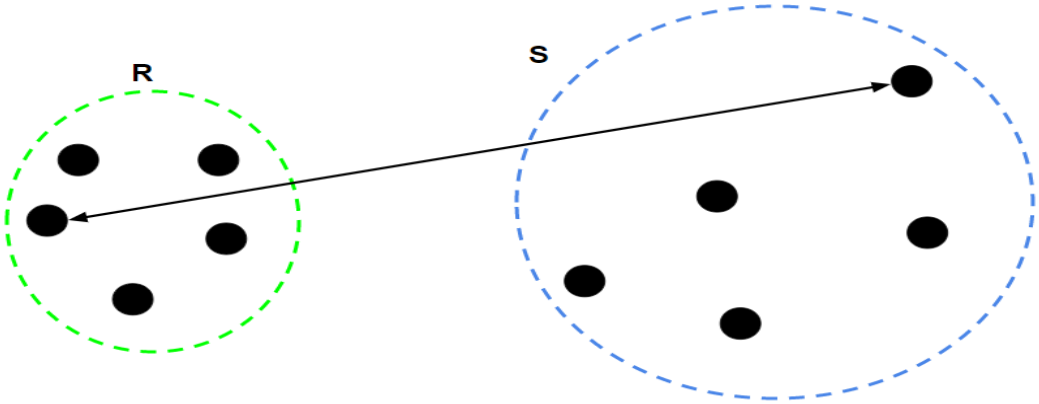
Q e) Explain the different linkages used in Hierarchical Clustering.

Ans e)

- **Single Linkage:** For two clusters R and S, the single **Linkage returns** the distance between 2 clusters is defined as the shortest distance between points in the two clusters

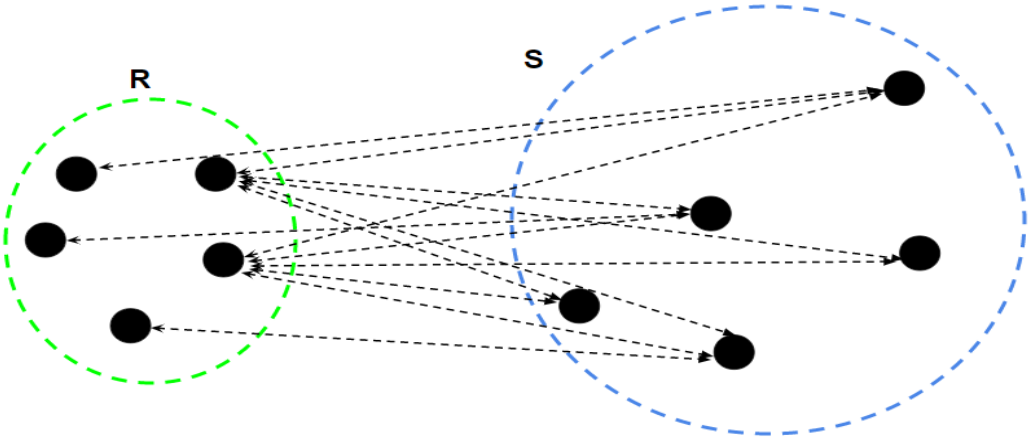


2. **Complete Linkage:** For two clusters R and S, the single linkage returns the maximum distance between any 2 points in the clusters



3. **Average Linkage:** Here, the distance between 2 clusters is defined as the average distance between every point of one cluster to every other point of the other cluster.

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.



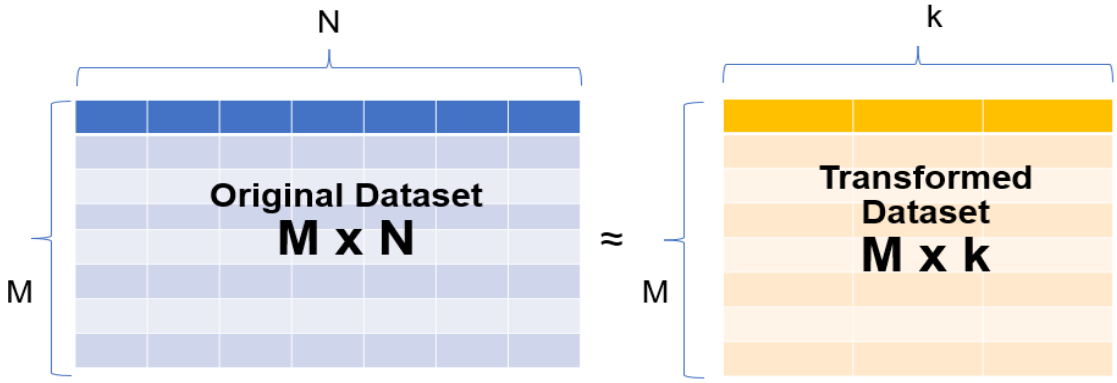
Ques a: Give at least three applications of using PCA.

Ans a: Application of PCA:-

I. PCA aims to orthogonally transform correlated variables to a smaller set of linearly uncorrelated variables (principal components). Hence, this will reduce the **multicollinearity thus leading to a better data.**

II. With PCA, you can transform 100 features into only 10 **relevant** new features (which can represent 90% characteristics of original 100 features). Using PCA features is more efficient.

III. The primary application of PCA is dimension reduction. If you have high dimensional data, PCA allows you to reduce the dimensionality of your data so the majority of the variation than exists in your data across many high dimensions is captured in fewer dimensions.



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

Ans: Basis: The **basic definition** of basis vectors is that they're a certain set of vectors whose linear combination is able to explain any other vector in that space.

A: 'basis' is a unit in which we express the vectors of a matrix.

B: Simple change of basis led to dimensionality reduction in the case of the roadmap example and then understood how you can represent the same data in multiple basis vectors.

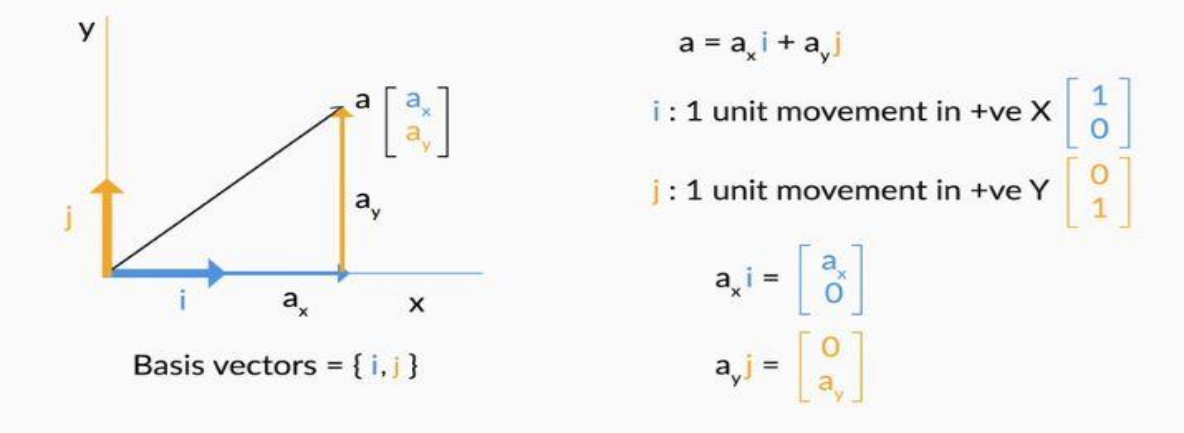
C: Any vector **'a'** (ax,ay) can be represented in a 2-D space, using the following notation :

$$a=axi+ayj$$

or

$$a=ax\cdot[10]+ay\cdot[01]$$

Visually, it can be represented as follows:



Variance is information:

Variance is a measure of spread, and indicates how far a set of number are spread out from their average value. for a one dimensional array X, the variance s2 is:

- Xi = The value of the I th entry of array X
- X bar = the average of X
- n = the number of entries

$$s^2 = \frac{\sum_i^n (X_i - \bar{X})^2}{(n-1)} = \frac{\sum_i^n (X_i - \bar{X})(X_i - \bar{X})}{(n-1)}$$

If a column has more variance, then this column will contain more information.

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

Ans: Shortcoming of using PCA:

1. 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.

2. 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.

3. PCA is focused on finding orthogonal projections of the dataset that contains the highest variance possible in order to 'find hidden LINEAR correlations' between variables of the dataset. This means that if you have some of the variables in your dataset that are linearly correlated, PCA can find directions that represents your data.

4. 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.