1. What is the difference between supervised and unsupervised learning? Give some examples to illustrate your point.
2. Mention a few unsupervised learning applications.
3. What are the three main types of clustering methods? Briefly describe the characteristics of each.
4. Explain how the k-means algorithm determines the consistency of clustering.
5. With a simple illustration, explain the key difference between the k-means and k-medoids algorithms.
6. What is a dendrogram, and how does it work? Explain how to do it.
7. What exactly is SSE? What role does it play in the k-means algorithm?
8. With a step-by-step algorithm, explain the k-means procedure.
9. In the sense of hierarchical clustering, define the terms single link and complete link.
10. How does the apriori concept aid in the reduction of measurement overhead in a business basket analysis? Give an example to demonstrate your point.

Q.1What is the difference between supervised and unsupervised learning? Give some examples to illustrate your point.

Ans: supervised Machine Learning: supervised Machine Learning is type of the machine learning in which we provide the labled data to the machine and then we use that data to train the machine learning model. Supervised learning model takes direct feedback to check if it is predicting correct output or not.

Unsupervised Learning:Unsupervised learning algorithms are trained using unlabeled data. Unsupervised learning model does not take any feedback.

Q. 2. Mention a few unsupervised learning applications.

Ans: clustering, visualization, dimensionality reduction, finding association rules, and anomaly detection

Q.3. What are the three main types of clustering methods? Briefly describe the characteristics of each. Ans: Three Main types of the clustring methods are: 1. Centroid-based Clustering 2. Distribution-based Clustering 3. Hierarchical Clustering 4. Density-based Clustering

1. Centroid-based clustering: Organizes the data into non-hierarchical clusters, in contrast to hierarchical clustering defined below. k-means is the most widely-used centroid-based clustering algorithm. Centroid-based algorithms are efficient but sensitive to initial conditions and outliers.

2. Distribution-based Clustering : This clustering approach assumes data is composed of distributions, such as Gaussian distributions. The distribution-based algorithm clusters data into three Gaussian distributions. As distance from the distribution's center increases, the probability that a point belongs to the distribution decreases.

3. Hierarchical Clustering: Hierarchical clustering creates a tree of clusters. Hierarchical clustering, not surprisingly, is well suited to hierarchical data. Another advantage is that any number of clusters can be chosen by cutting the tree at the right level.

4. Density-based Clustering: Density-based clustering connects areas of high example density into clusters. This allows for arbitrary-shaped distributions as long as dense areas can be connected. These algorithms have difficulty with data of varying densities and high dimensions. Further, by design, these algorithms do not assign outliers to clusters.

Q. 4. Explain how the k-means algorithm determines the consistency of clustering.

Sol: K-means is perfect for getting to know your data and providing insights on almost all datatypes. Whether it is an image, a figure or a piece of text, K-means is so flexible it can take almost everything. K-means uses an iterative refinement method to produce its final clustering based on the number of clusters defined by the user (represented by the variable K) and the dataset. For example, if you set K equal to 3 then your dataset will be grouped in 3 clusters, if you set K equal to 4 you will group the data in 4 clusters, and so on.

When we define the value of K we are actually telling the algorithm how many means or centroids you want (if you set K=3 you create 3 means or centroids, which accounts for 3 clusters). A centroid is a data point that represents the center of the cluster (the mean), and it might not necessarily be a member of the dataset.

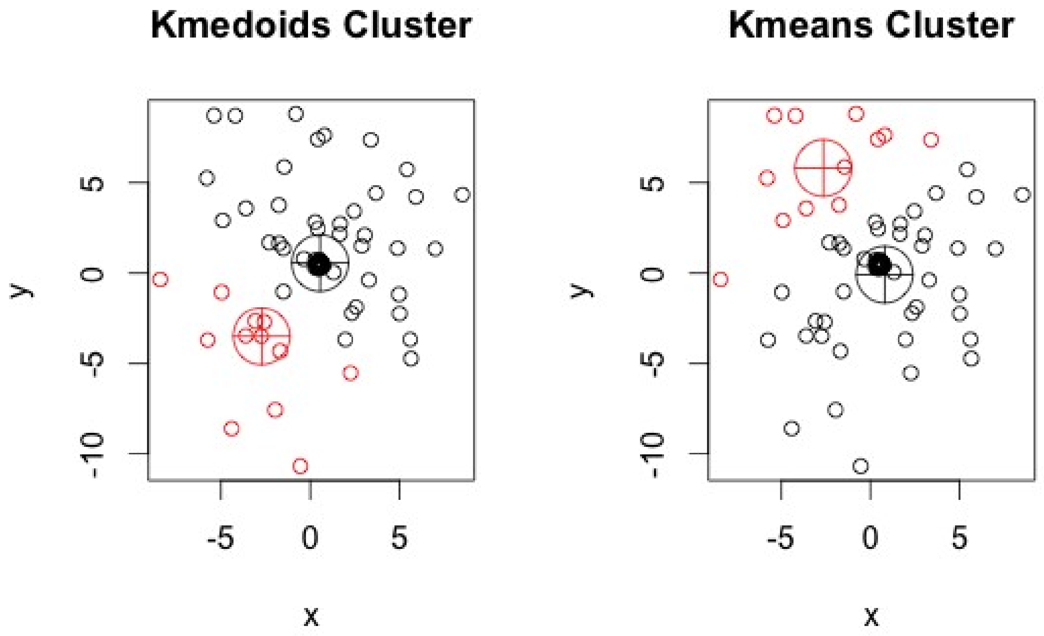
This is how the algorithm works:

K centroids are created randomly (based on the predefined value of K) K-means allocates every data point in the dataset to the nearest centroid (minimizing Euclidean distances between them), meaning that a data point is considered to be in a particular cluster if it is closer to that cluster’s centroid than any other centroid Then K-means recalculates the centroids by taking the mean of all data points assigned to that centroid’s cluster, hence reducing the total intra-cluster variance in relation to the previous step. The “means” in the K-means refers to averaging the data and finding the new centroid

"""The algorithm iterates between steps 2 and 3 until some criteria is met (e.g. the sum of distances between the data points and their corresponding centroid is minimized, a maximum number of iterations is reached, no changes in centroids value or no data points change clusters)""

Q. 5. With a simple illustration, explain the key difference between the k-means and k-medoids algorithms.

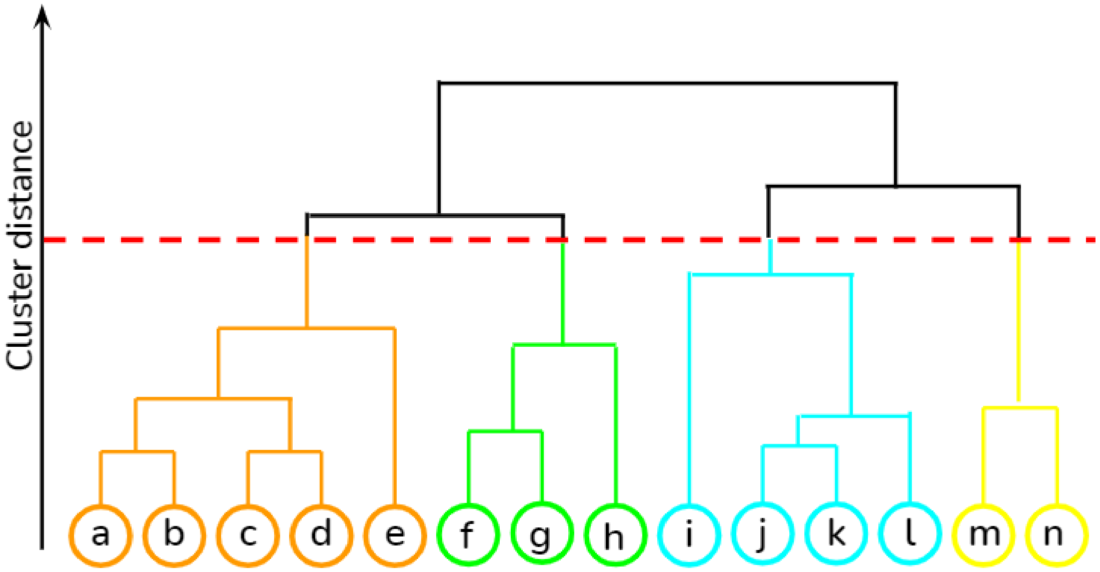
Ans: K-means attempts to minimize the total squared error, while k-medoids minimizes the sum of dissimilarities between points labeled to be in a cluster and a point designated as the center of that cluster. In contrast to the k -means algorithm, k -medoids chooses datapoints as centers (medoids or exemplars).



Q.6. What is a dendrogram, and how does it work? Explain how to do it.

Ans: Dendrogram The sole concept of hierarchical clustering lies in just the construction and analysis of a dendrogram. A dendrogram is a tree-like structure that explains the relationship between all the data points in the system.

However, like a regular family tree, a dendrogram need not branch out at regular intervals from top to bottom as the vertical direction (y-axis) in it represents the distance between clusters in some metric. As you keep going down in a path, you keep breaking the clusters into smaller and smaller units until your granularity level reaches the data sample. In the vice versa situation, when you traverse in up direction, at each level, you are subsuming smaller clusters into larger ones till the point you reach the entire system.



Q. 7. What exactly is SSE? What role does it play in the k-means algorithm?

Ans: SSE is defined as the sum of the squared distance between centroid and each member of the cluster. It help to chose the number of clusters using the elbow method.

Q. 8. With a step-by-step algorithm, explain the k-means procedure.

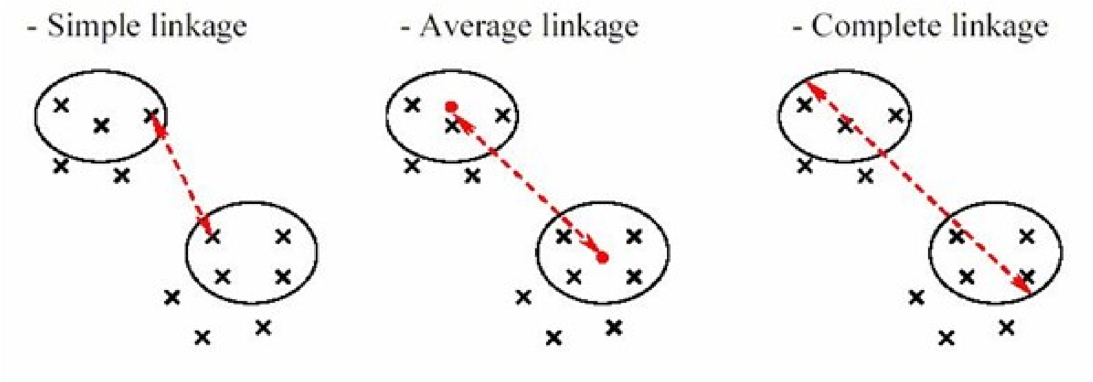
This is how the k-means algorithm works:

1. K centroids are created randomly (based on the predefined value of K)
2. K-means allocates every data point in the dataset to the nearest centroid (minimizing Euclidean distances between them), meaning that a data point is considered to be in a particular cluster if it is closer to that cluster's centroid than any other centroid
3. Then K-means recalculates the centroids by taking the mean of all data points assigned to that centroid's cluster, hence reducing the total intra-cluster variance in relation to the previous step. The "means" in the K-means refers to averaging the data and finding the new centroid
4. The algorithm iterates between steps 2 and 3 until some criteria is met (e.g. the sum of distances between the data points and their corresponding centroid is minimized, a maximum number of iterations is reached, no changes in centroids value or no data points change clusters)

Q. 9. In the sense of hierarchical clustering, define the terms single link and complete link.

Single-link: clusters at step are maximal sets of points that are linked via at least one link (a single link) of similarity.

Complete-link clusters at step are maximal sets of points that are completely linked with each other via links of similarity .


- Simple linkage
- Average linkage
- Complete linkage