Q1. What is the most appropriate no. of clusters for the data points represented by the following dendrogram:

Answer: Option (b) 4

Q2. In which of the following cases will K-Means clustering fail to give good results?

1. Data points with outliers
2. Data points with different densities
3. Data points with round shapes
4. Data points with non-convex shapes

Answer: Option d) 1, 2 and 4

Q3. The most important part of_____ is selecting the variables on which clustering is based.

a) interpreting and profiling clusters
b) selecting a clustering procedure
c) assessing the validity of clustering
d) formulating the clustering problem

Answer: Option d)

Q4. The most commonly used measure of similarity is the____ or it's square.

a) Euclidean distance
b) city-block distance
c) Chebyshev's distance
d) Manhattan distance

Answer: Option a)

Q5. ___a clustering procedure where all objects start out in one giant cluster. Clusters are formed by dividing this cluster into smaller and smaller clusters.

a) Non-hierarchical clustering
b) Divisive clustering
c) Agglomerative clustering
d) K-means clustering

Answer: Option b)

Q6. Which of the following is required by K-means clustering?

a) Defined distance metric
b) Number of clusters
c) Initial guess as to cluster centroids
d) All answers are correct

Answer: Option d)

Q7. The goal of clustering is to-

a) Divide the data points into groups
b) Classify the data point into different classes
c) Predict the output values of input data points
d) All of the above

Answer: Option d)

Q8. Clustering is a

Answer: Unsupervised learning

Q9. Which of the following clustering algorithms suffers from the problem of convergence at local optima?

a) K- Means clustering
b) Hierarchical clustering
c) Diverse clustering
d) All of the above

Answer: option a)

Q10. Which version of the clustering algorithm is most sensitive to outliers?

a) K-means clustering algorithm
b) K-modes clustering algorithm
c) K-medians clustering algorithm
d) None

Answer: option a)

Q11. Which of the following is a bad characteristic of a dataset for clustering analysis-

a) Data points with outliers
b) Data points with different densities
c) Data points with non-convex shapes
d) All of the above

Answer: option d)

Q12. For clustering, we do not require-
a) Labeled data
b) Unlabeled data
c) Numerical data
d) Categorical data

Answer: option a)

Q13. How is cluster analysis calculated?

Answer: Cluster analysis is an unsupervised learning algorithm. Different Algo's have different hyperparameters.

Clustering any point together, the main metric is distance metric, without selecting the correct distance metric w.r.t our analysis. Deriving relevant KPIs and feature selection are the most important criteria for clustering analysis.

Q14. How's cluster quality measured?
Answer: The very basic things we can think of to measure cluster quality is distance between the clusters and within the clusters. Jargons for prior statements are inter and intra cluster.

Following are the few methods to measure clustering quality.
Dunn Index: $D = \max d(i,j) / \min d'(k)$

Where, $d(i,j)$= distance between cluster $c(i)$ and $c(j)$

$d'(k)$= intra cluster distance within cluster k. (Basically farthest distance within cluster k)

By equation, we can see that if the numerator is high means distance between clusters is high => D will be high => very good clustering.

Ward's method: Similarity of two clusters is based on the increase in squared error when two clusters are merged. (Used in agglomerative clustering)

Elbow Method: Is a heuristic way of plotting the number of clusters w.r.t their quality. While observing the graph we can see elbow structure or global minima, which gives us the total clusters where quality is best or in other words distance between clusters are max.

Q15. What is cluster analysis and its types?
Ans. Cluster analysis is performed when the labels are not available, its an unsupervised learning algorithm.

Types of clustering algo's:
K-mean clustering, K-means++, K-medoids, Agglomerative, Divisive and most popular one DBSCAN(Density Based spatial clustering of application with noise)