Show your working in all calculations.

2. This question is related to **Clustering**.

a) Describe briefly the main objective of Cluster Analysis.

b) Explain the difference between *partitional clustering* amd *hierarchical clustering* approaches.

c) Describe briefly and illustrate graphically the *elbow method* for selecting the number of clusters in k-means clustering.

d) Discuss the basic DIFFERENCE between the agglomerative and divisive hierarchical clustering algorithms. Explain also the difference between single (MIN) and complete (MAX) linkage in the context of hierarchical clustering.

Cluster analysis groups data objects based on the information in the data that describes the objects and their relationships. Its goal is that the objects within a group be similar to each other and different to objects within other groups. The greater the similarity (homogeneity) within a group and the greater the difference between groups, the better or more distinct the clustering.

Partitional clustering divides data objects into non-overlapping subsets. Each data point is in exactly one subset.
Hierarchical clustering have nested clusters, organized as hierarchical tree. Each clusters can have sub-clusters. Each node in the tree (except for the leaf node) is a union of its children node. The root node is the cluster of all data objects.

K-means clustering is a partitional clustering approach where each cluster are associated with a centroid (centre point), each point is associated to the cluster with the closest centroid. Number of clusters (K) must be specified.
Ways of selecting initial centroid: randomly with multiple points and choose the one with least error, non-random (e.g. the most diverse mutually distant k-subset of the data as initial cluster centres.
Pre-processing step of K-means Normalize data, Eliminate outliers
Post-processing step: Eliminate clusters with relatively small points (that may represent outliers), Split loose clusters with high SSE, merge clusters that are close and have relatively low SSE.
Elbow method is a means of selecting the K (number of clusters) in advance. It involves iteratively increasing the number of clusters and observing the SSE gains (which tends to flatten after some iterations), the K selected is the one that has a reasonable trade-off between model complexity(number of clusters) and SSE.
Limitations of K-Means: Size, density, non-globular shapes and outliers
Overcoming K-Means Limitations: Find many small clusters and merge together.

Agglomerative clustering algorithms starts with each point as individual clusters and merges closest pair of clusters until a single cluster is (or k clusters are) obtained.
Divisive Hierarchical clustering algorithms start with a single all-inclusive cluster and at each step, splits a cluster until each cluster contains a single point (or there are k clusters
Strengths of Hierarchical Clustering: Any number of clusters can be obtained by cutting the dendogram at each point.
MIN, MAX, Group Average and Distance between centroids are ways of defining inter-cluster similarity. MIN uses the two most closest (similar) points to determine similarities between two clusters. It can handle non-globular shapes but sensitive to noise and outliers
MAX uses the least similar (most distant) points to define similarity between two clusters. It is biased towards globular shapes and less sensitive to noise and outliers.

DBSCAN is a Density based clustering algorithm that partition points into dense regions separated by not-so-dense regions. How is density measured: Density at a point: is defined as the number of points within a ball of radius Eps.
Dense region is a region with at least MinPts points within a ball of radius Eps.
A core point is a point with at least MinPts points within a distance Eps. They are in a dense region and belong to the interior of the cluster.
A border point has less than MinPts points within a distance Eps but is in the Eps-neighbourhood of at least one core point.
A noise point is neither a core point nor a border point. A point is density-connected to another if there is path of edges between them (edge is formed by two core points within distance Eps from each other).
DBSCAN algorithm starts by labelling core, border and noise points, eliminate the noise points, connect core points that are within Eps distance to form edges, make each group of connected core points form a cluster, assign each border point to one of its associated core points cluster.