```r
# We aim to investigate the similarities among various soil pollutants by
# analyzing the chemical formulae of their main component.
# One of your collaborators has devised a metric distance measure for chemical
# formulae based on classic distances between graph objects.
# The file molecules.txt contains the matrix of the pairwise distances between
# 90 different molecules of soil pollutants.
# In order to explore the similarities between these molecules, we will employ
# a cluster analysis approach.
# a) Which clustering methods discussed in class are suitable for this case?
#    Provide a precise justification for your answer.
# b) Perform hierarchical clustering of the molecules using average linkage.
#    Report the dendrogram. Determine an appropriate number of clusters and cut
#    the dendrogram accordingly. Report the sizes of the resulting clusters.
# c) Now, let us explore the DBSCAN approach.
#    Using a minPts value of 3 and choosing a consistent value for eps (with an
#    accuracy of .05), run the DBSCAN algorithm.
#    Justify your choice for eps and report any computations or plots involved
#    in this choice. Also, provide the number of clusters discovered and their
#    respective sizes. Is the result satisfactory?
# d) Run DBSCAN again, this time with minPts = 10 and eps = 0.15.
#    Report the number of clusters identified and their sizes.
#    Suggest a quantitative method for comparing the quality of the clustering
#    results obtained from this DBSCAN run and the hierarchical clustering
#    conducted in b).
#    Based on this method, select the best clustering procedure.
# e) Is there a way to visualize the molecules in a two-dimensional plot?
#    Report the plot, showing through it the results of the chosen clustering
#    procedure from the previous question. Assess whether the plot tends to
#    underestimate or overestimate the true distances.

# question a)
# We can use hierarchical agglomerative clustering, k-medoids, DBSCAN.
# We can't use k-means because we have a distance matrix and not a dataset,
# so we can't compute means. We can use however k-medoids, which is a
# k-means with the restriction that the centroids must be one of the points
# in the dataset (in this case, these are called medoids).
# The same goes for ward linkage method: we can't use it because we have a
# distance matrix and not a dataset, so we can't compute the
# ESS_j = sum x in C_j (d(x, c_j))^2, with c_j = mean of C_j.

# question b)
molecules <- read.table("molecules.txt", header = TRUE)
typeof(molecules)
dim(molecules)
head(molecules)

molecules.dist <- as.dist(molecules)
typeof(molecules.dist)

molecules.dist.matrix <- as.matrix(molecules.dist, nrow = 90, ncol = 90)
typeof(molecules.dist.matrix)

image(1:90, 1:90, molecules.dist.matrix,
    xlab = "Molecule", ylab = "Molecule",
    main = "Molecules distance matrix"
)

molecules.hclust <- hclust(molecules.dist, method = "average")
# plot the dendrogram
svg("molecules_dendrogram.svg", width = 6, height = 6)
plot(
    molecules.hclust,
    main = "Molecules dendrogram, average linkage",
    hang = -0.1,
    labels = FALSE,
    xlab = "Molecule",
```

```
    cex = 0.6,
    sub = ""
)
k <- 3
rect.hclust(
    molecules.hclust,
    k = k,
    border = "red"
)
dev.off()

# cut the dendrogram for k = 3 clusters
cluster.average <- cutree(molecules.hclust, k = k)

# interpret the clusters
table(cluster.average)

# Compute silhouette scores
library(cluster)

sil.average <- silhouette(cluster.average, dist = molecules.dist.matrix)
summary(sil.average)

# question c)
library(dbscan)

minPts <- 3
svg("molecules_eps.svg", width = 5, height = 5)
kNNdistplot(molecules.dist, k = minPts)
abline(h = 0.102)
dev.off()

eps <- 0.105
dbs1 <- dbscan(molecules.dist, eps = eps, minPts = minPts)
dbs1

# Compute silhouette scores:
clustered.indexes <- which(dbs1$cluster != 0)
length(clustered.indexes)
molecules.clustered <- molecules.dist.matrix[clustered.indexes, clustered.indexes]

labels <- dbs1$cluster[clustered.indexes]
sil1.dbscan <- silhouette(labels, dist = molecules.clustered)
summary(sil1.dbscan)
# We can see for the first cluster that the silhouette score is ~ 0.4: we can try to
# improve the clustering by changing the parameters.

# question d)
minPts <- 10
kNNdistplot(molecules.dist, k = minPts)
abline(h = 0.25)
eps <- 0.15

dbs2 <- dbscan(molecules.dist, eps = eps, minPts = minPts)
dbs2

clustered.indexes <- which(dbs2$cluster != 0)
molecules.clustered <- molecules.dist.matrix[clustered.indexes, clustered.indexes]

labels <- dbs2$cluster[clustered.indexes]
sil2.dbscan <- silhouette(labels, dist = molecules.clustered)
summary(sil2.dbscan)
```