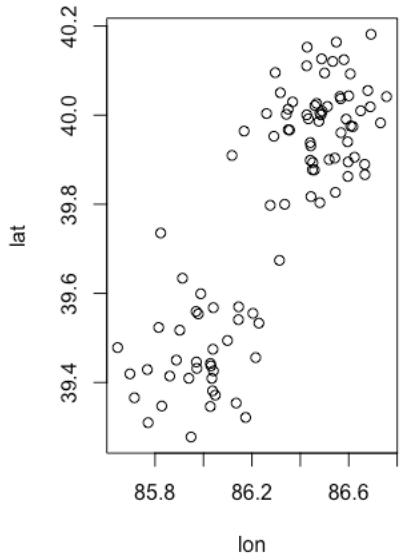
APPLIED STATISTICS EXAM

DATE: 12/07/2022
STUDENT: FILIPPO CIPRIANI
PERSONA CODE: 10956877

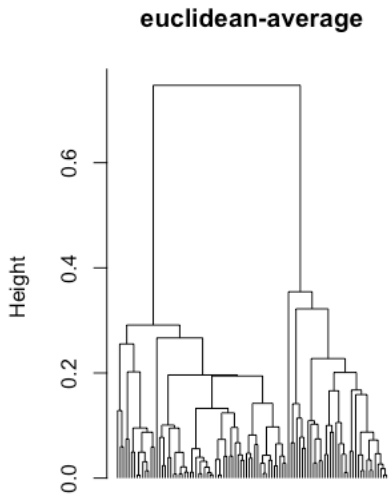
EXERCISE NUMBER 2

POINT A)

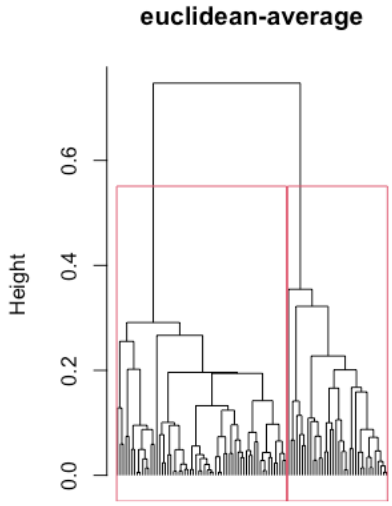
We take a look at the data and see that maybe there are 2 clusters of data.



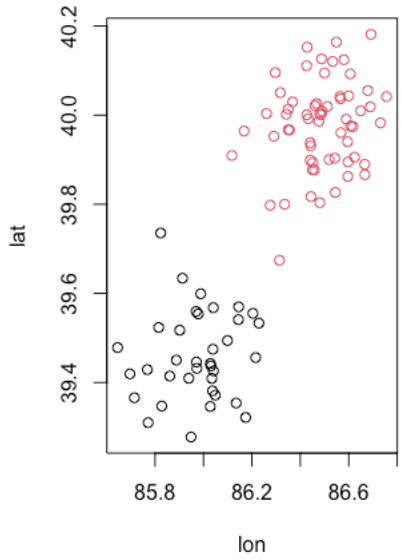
We compute the distance matrix for the data using Euclidean distance, then using a average linkage we build the dendrogram here:



Given the previous observation and the dendogram height it's clear that the optimal number of clusters is 2.



The clustering is not bad, we have a cophenetic coefficient of 0.8930955 and visualizing the datas we can see that it performs well in separing the clusters we have seen before:



We have two clusters of sizes:

36 - 61

POINT B)

We check that the datas belonging to a certain cluster are normal bivariate by performing a shapiro test for each cluster of data, we have pvalues:

Clust1: 0.6228

Clust2: 0.8584

P- values high enough not to reject the hypothesis of normality. We perform a Bartlett test for the same covariance structures assumptions between clusters, obtaining a high p-value of 0.5642 and therefore verifying all the MANOVA assumptions.

We perform a Manova using the cluster labels: the Wilks statistic suggests us that there's a sensible difference between the groups, the pvalue is basically zero, 2.2e-16 . So there is statistical evidence that membership to a cluster rather than another makes the difference on the mean positions.

The model of the MANOVA is:

Model: X.ij = mu + tau.i + eps.ij; eps.ij~N_2(0,Sigma), X.ij, mu, tau.i in R^2

Where tau.i i=1,2 is the effect of the clustering membership. The coefficients of the MANOVA model are reported below:

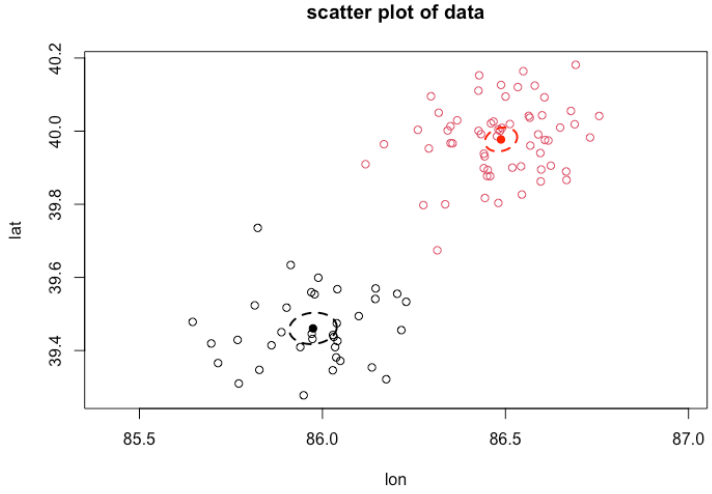
	lon	lat
(Intercept)	85.9741972	39.4606028
Cluster2	0.5135979	0.5166185

POINT C)

The clusters centroids are:

	lon	lat
Clust1	85.9742	39.46060
Clust2	86.4878	39.97722

And below is the scatter plot of the data, with a confidence region at level 0.95 for the mean of each cluster:



The equation for each of the conf. region is:

{ m \in R^2 s.t. n * (x.mean-m)' %*% (x.cov)^-1 %*% (x.mean-m) < cfr.fisher }

Where x.mean,n,x.cov and cfr.fisher = ((n-1)*p/(n-p))*qf(1-alpha,p,n-p), alpha = 0.05 have been calculated for each cluster.