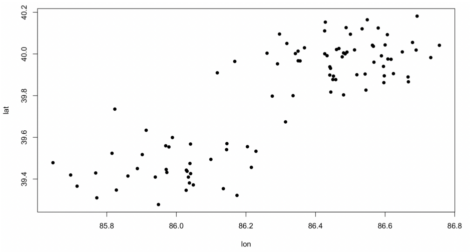
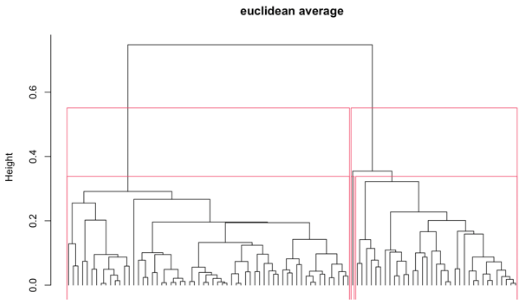
Problem 2

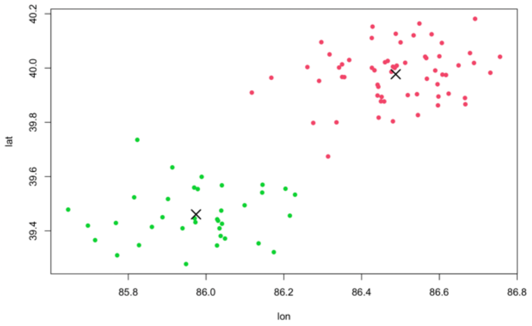
9.5/10

We look at our data and suppose that there are 2 clusters.



The we compute the distance matrix using Euclidean distance and using average linkage. We build the dendrogram. The cophenetic coefficient is: 0.8930955 so we expect a not noisy dendrogram





As we can see the best number of clusters is 2 and the scatterplot of data is coherent.

We check the assumption for the MANOVA model:

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

X.ij, mu, tau.i in R\^p

Test: H0: tau.1 = tau.2 = (0,0) H1: (H0)\^c

Multivariate normality in each group is satisfied since we get the following p-values 0.8388 0.6132 from Shapiro test. The same covariance structure in each group too by visual inspection seems ok and the Bartlett test confirm with a p-value of 0.5642 so we can't reject the null hypothesis of equality between the variances.

We perform the MANOVA test using as group the cluster label and the Wilks statistic suggests us that there's a sensible difference between the mean of the groups. The p-value is nearly 0 (< 2.2*10^(-16)) so there is difference between the mean of the groups. By ANOVA test we see that both lat and long are different.

The estimated parameters are the following:

	lon	lat
(Intercept)	86.4877951	39.9772213
groups2	-0.5135979	-0.5166185

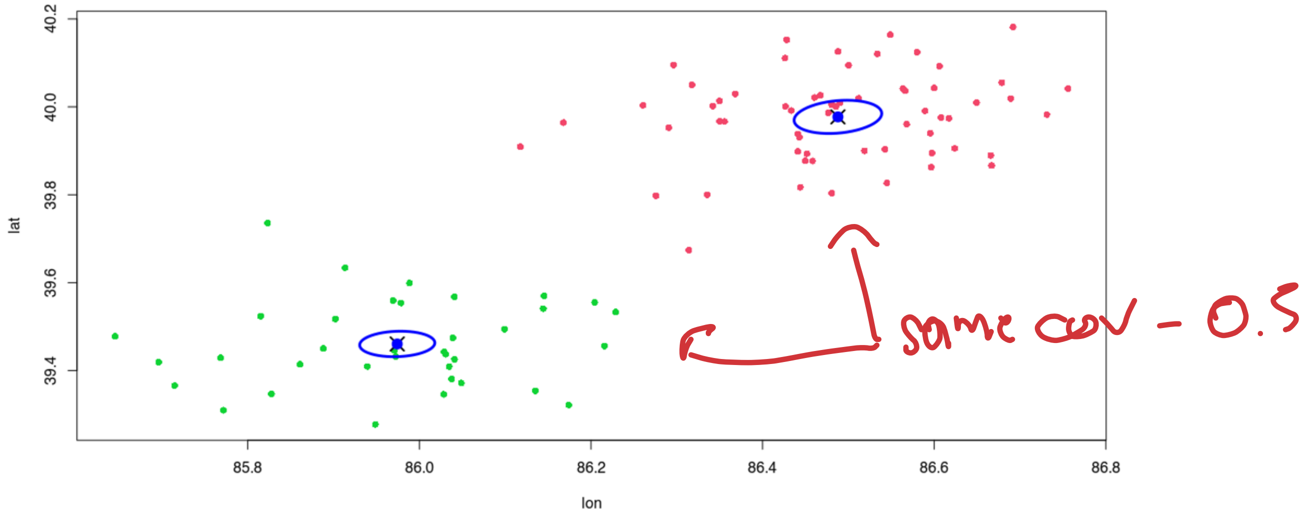
sigma = 0.1217148

this shall be a matrix 2x2

-0.5

The center of the clusters are: (86.4878, 39.97722) and (85.9742, 39.46060) and their dimension are: 61 36.

Calculating confidence regions at level 95% for the means we get the following result.



same cov - 0.5