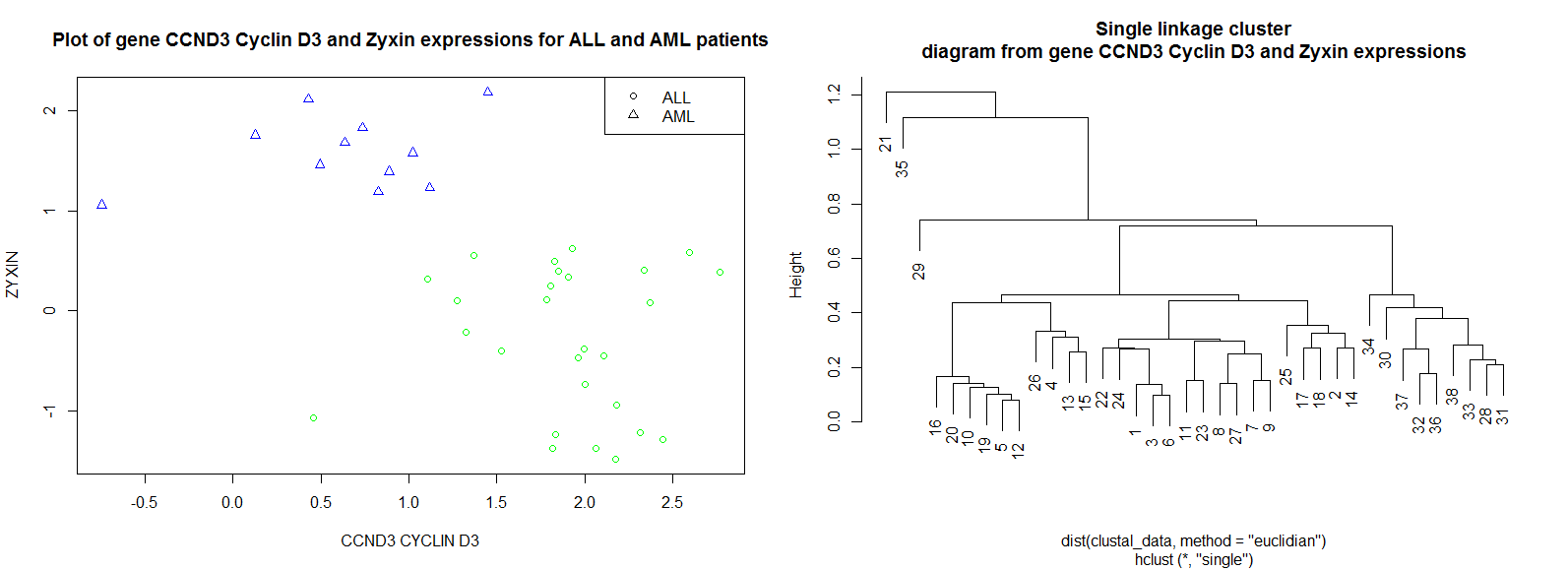
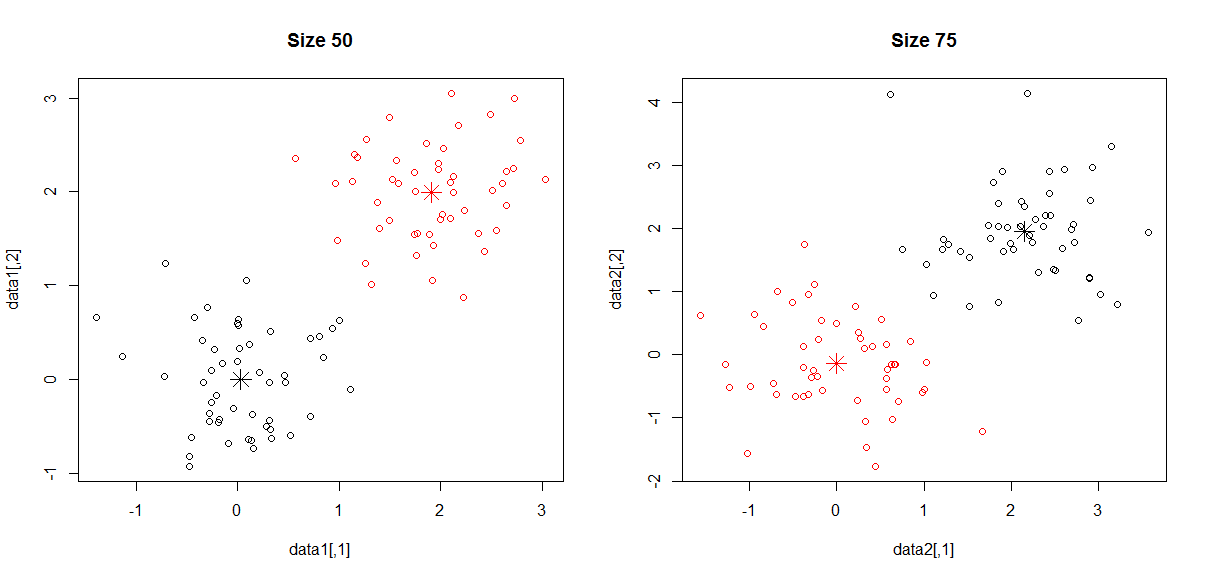
Nichollette Acosta

ITCS 3162 | Term Project

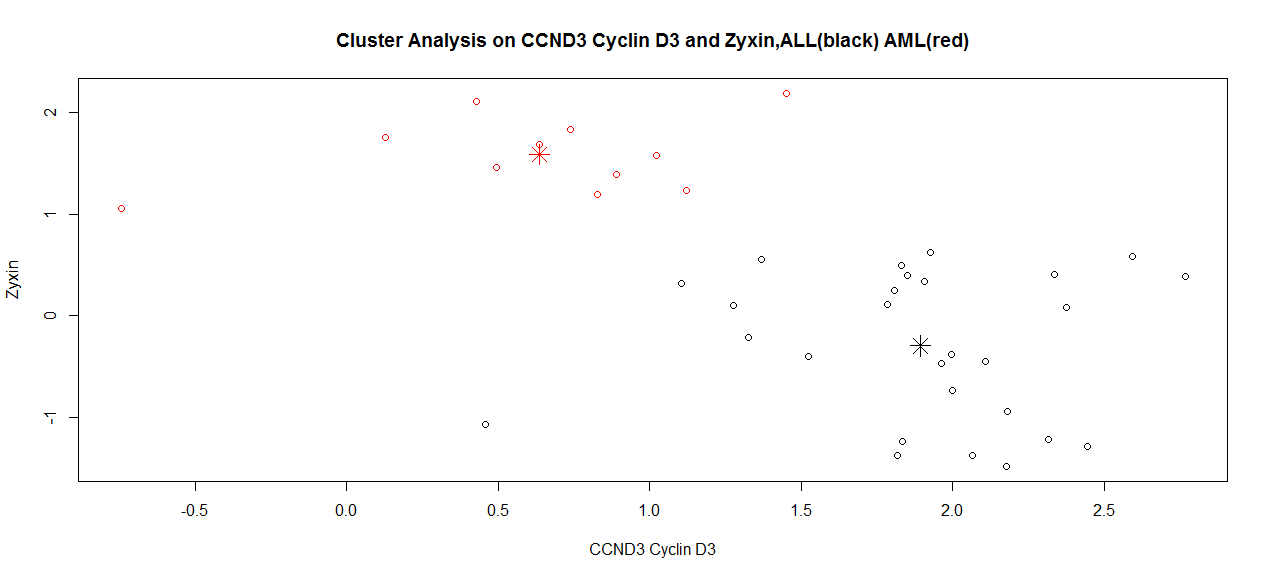
Gene expression data collected by Golub et al. (1999) are among the classical bioinformatics. There is a selection of the set called golub and is contained in the multtest package, a part of Bioconductor. The data consistof gene expression values of 3051 genes (rows) from 38 leukemia patients. Twenty-seven patients are diagnosed as acute lymphoblastic leukemia (ALL) and eleven as acute myeloid leukemia (AML). We concentrate on expression values of the gene, “CCND3 Cyclin D3.” We simulated gene expressions from two normal populations, N(0,0.5) and N(2,0.5).On the total of 100 data points on a k-means cluster,of 2-means, analysis is performed. The output of the k-means cluster led us to hypothesis of the populations being, (0,0) and (2,2). We then used a bootstrap to estimate a 95% confidence interval around cluster means. We re-sampled with replacement 1000 times and produced the corresponding quantiles for the confidence intervals. The bootstrap confidence intervals accept the null hypothesis that the cluster population means are (0,0) and (2,2). From this we can see that the expression values of “CCND3 Cyclin D3” and “Zyxin” are closely related to the distinction between AML and ALL. These two clusters discriminate the ALL patients from the AML patients. This can also be seen from Figure 3, where expression values of CCND3 Cyclin D3 are depicted on the horizontal axis and those of Zyxin on the vertical, and the ALL patients are in red and the AML patients in black. By the bootstrap the cluster means and their confidence intervals can be estimated.



**Figure A**



**Figure B: Unsupervised K-means Analysis**

****

**Figure C**