INTRODUCTION TO BIOINFORMATICS
ASSIGNMENT #4

I have used TM4 MEV tool in the analysis of the microarray dataset. First, I uploaded the data set that is given. Then I have applied k-means clustering algorithm to this data set since I know the k that is two as healthy and disease tissues. Then I've answered the questions.

1) Clustering algorithm gives the result in the below:
   n = 17 and the samples in this cluster (let us say healthy ones) are: sample29, sample28, sample1, sample5, sample4, sample8, sample7, sample30, sample11, sample20, sample12, sample22, sample23, sample24, sample25, sample26, sample27

   m = 13 and the samples in this cluster are (let us say diseased ones): sample19, sample18, sample17, sample21, sample3, sample10, sample2, sample15, sample9, sample16, sample13, sample14, sample6

2) I wrote a code to complete this part. After getting separated the samples as diseased and healthy tissues, I summed all of the samples by negating the healthy ones. In addition to this, when summing them up, I divided the healthy ones by 13, and diseased ones by 17 to minimize the effect of having different size of samples. I have also looked the GeneMAD and GeneSD analysis on the MEV tool. They're all compatible with each other with little differences. According to these(mostly GeneMAD) analysis, the results are below:
   10 genes that are highly expressed in diseased tissues are:
   205725_at, 220542_s_at, 204892_x_at, 203021_at, 211296_x_at, 207783_x_at, 210646_x_at, 206559_x_at, 213477_x_at, 212869_x_at
   10 genes that are highly expressed in healthy tissues are:
   215691_x_at, 222229_x_at, 207761_s_at, 215299_x_at, 207169_x_at, 200803_s_at, 217109_at, 221651_x_at, 213048_s_at, 214359_s_at