Stat 407 Lab 7 Cluster Analysis Fall 2012
Due: Wednesday, Oct 3 in class Hand in one solution per group.

Purpose: This lab is about cluster analysis. We'll look at Dr Cook's music data and cluster the music tracks into groups according to similarity based on variables calculated on the first 40 seconds of sound. There are 62 data points (tracks), 11 tracks from Abba, Beatles, 10 tracks from Eels, 8 tracks from Beethoven 6 tracks from Mozart, 13 tracks from Vivaldi and 3 tracks from Enya. The variables are:

LVar, LAve, LMax — the average, variance, maximum of the frequencies of the left channels

LFEner — an indicator of the amplitude or loudness of the sound.

LFreq — an average of the most common frequencies computed using the `periodogram` function.

1. Write a paragraph describing how you would expect the music clips to cluster.
2. Standardize the variables. You'll need to use these standardized variables for the cluster analysis. Explain why this is necessary.
3. Calculate the interpoint distance matrix. using Euclidean distance. Why is the length of the result of the dist function on the music data equal to1891? Which clips are closer together "Knowing Me" and "Dancing Queen" or "Knowing Me" and "Yesterday"?
4. Compute a hierarchical cluster analysis with single linkage and also Ward's linkage method. Report the dendrograms. Which linkage method do you think gives better results? Explain your reasoning.
5. Using your preferred method, based on the dendrogram and your own intuition about the music, how many clusters would you make for this data? Think about using the clusters as becoming a play list on your iPod or mp3 player, so that a fairly small number but similar types of music will be played consecutively. Report the clusters, using the name of the track.

Useful R code:

- Reading data

```
> music <- read.csv(file.choose(), row.names=1) # music-plusnew-sub.csv
```

- To standardize a variable

```
> library(reshape)
> music.std <- rescaler(music)
```

- Calculate distances, without the artist and type of music variables

```
> ?dist
> music.dist <- dist(music.std[,-c(1,2)])
> length(music.dist)
> print(music.dist)
```

- To do a hierarchical cluster analysis use the function `hclust`.

```
> ?hclust
> music.hc <- hclust(music.dist, method="single")
> plot(music.hc, hang=-1)
> music.hc <- hclust(music.dist, method="ward")
> plot(music.hc, hang=-1)
> cl <- cutree(music.hc, 5)
> music[cl==1, 1:2]
> music[cl==2, 1:2]
> music[cl==3, 1:2]
> music[cl==4, 1:2]
> music[cl==5, 1:2]
> music$cl <- as.factor(cl)
> library(GGally)
> ggparcoord(music, columns=3:7, groupColumn="cl")
```