Stat 407 Homework 4 Fall 2012
Due: Wednesday, November 14 in class Individual work is expected.

In this homework you will look at data collected on movies from the IMDB web site. There is a subset of action or romance movies released between 1986 and 2004. The variables of interest are:

budget	how much it cost in dollars
length	in minutes
rating	average consumer rating
votes	number of people rating the movies
Action	is 1 if it was an action movie, and 0 if it was a romance movie

The training data is in the file `movies-train.csv`. The test data is in `movies-test.csv`.

1. Read in the data, and descriptively summarize the data. Make a scatterplot matrix, colored by movie genre. Compute the means, standard deviations and correlation matrix for each movie genre.
2. Check the normality of the four variables. And use transformations to make it reasonably normally distributed. Report what you have done to achieve this.
3. Build a linear discriminant analysis classification on the training data to determine if it is possible to classify movies as either Action or Romance based on the budget, length, rating and votes. Report the LDA rule, and the training error. Report the test error.
4. Using the training data, conduct a MANOVA test on the means for the types of movies. Write down what hypothesis is being tested. Summarize the results, the test statistic value, and the p-value of the test.
5. Using ANOVA, determine which of the variables are the most important contributors to the significant difference between the means. Write a paragraph describing how Action movies differ from Romance movies, on average.

Sample code:

- Reading data

```
m.tr<-read.csv(file.choose()) # movies-train.csv
m.ts<-read.csv(file.choose()) # movies-test.csv
m.tr$Action <- factor(m.tr$Action)
m.ts$Action <- factor(m.ts$Action)
```

- Plotting

```
library(ggplot2)
library(GGally)
ggpairs(m.tr, columns=3:6, color="Action", upper="blank", alpha=0.2)
qplot(budget,data=m, geom="histogram")
qplot(budget,data=m, geom="histogram") + scale_x_log10()
qplot(budget,data=m, geom="histogram") + scale_x_sqrt()
```

- Summary

  ```
  colMeans(m[m$Action==1,c(4,5,9,10)])
  colMeans(m[m$Action==0,c(4,5,9,10)])
  var(m[m$Action==1,c(4,5,9,10)])
  var(m[m$Action==0,c(4,5,9,10)])
  ```

- LDA

  ```
  library(MASS)
  m.lda <- lda(Action~tbudget+length+rating+tvotes, data=m.tr, prior=c(0.5,0.5))
  table(m.tr$Action, predict(m.lda, m.tr)$class)
  ```

- MANOVA

  ```
  summary(manova(cbind(tbudget,length,rating,tvotes)~Action, m.tr), test="Wilks")
  summary(aov(tbudget~Action, m.tr))
  ```