Stat 407 Lab 5 Principal Component Analysis Fall 2012
Due: Wednesday, Sep 19 in class Hand in one solution per group.

Purpose: In this lab we will learn how to conduct principal component analysis. We will use PCA to assess livability in small American cities.

Conduct a principal component analysis on the ten rating variables of the livability in American small cities data. The data comes from a book that was published in the late 90s, on livability of small cities in the USA. The Des Moines Register featured the book, because Ames was ranked the number 2 best small city in which to live! The study used 10 ratings variables – Climate, Diversions, Economic, Education, Community, Health, House, Safety, Transportation, Urban – with each city getting a rating between 0-100 on each of these. The scores were combined to give an overall rating for each city, Score.

1. Load the data into ggobi. Check for outliers, clustering or nonlinear relationships. Comment on anything that you think might affect the principal component analysis. Also, find Ames (it has the second highest value on Score) and explain how Ames rates on the 10 rating variables.
2. All the scores are measured on a scale of 0-100, so why is it still necessary to use the correlation matrix, or standardize the data, before doing PCA? (Hint: Compute some summary statistics or make some plots.)
3. Write down the summary of the PCA, the table of eigenvectors, eigenvalues (variance), and cumulative percent of total variance, in the format used in class notes. (Be sure to make your output readable, eg rounding digits appropriately.)
4. Interpret the first principal component.
5. Compare the scores for the first principal component with the Score variable in the data (this is the rating the book gives for each city). Which city would be rated first using the Score variable, and which city by PCA? (You could make a plot of the Score variable against PC1, and compute the correlation between the two variables.) Explain why the two approaches might give cities different scores.

Notes:

- When you read the data in be sure to use the option `row.names=1` on the `read.csv` command. This will make it possible to see the town names corresponding to each case.
- Use the `prcomp` command to compute the principal components. Options `scale=T` standardizes the data, uses the correlation matrix as the input to PCA, and `retx=T` returns the principal component scores as part of the result.